Is it time to use machine learning survival algorithms for survival and risk factors prediction instead of Cox proportional hazard regression? A comparative population-based study

Short title: Machine learning survival algorithms

Sara Morsy[1,2,3], Truong Hong Hieu[3,4⊕], Abdelrahman M Makram[3,5⊕], Osama Gamal Hassan[3,6], Nguyen Tran Minh Duc[3,4], Ahmad Helmy Zayan[3,7], Le-Dong Nhat-Nam[3,8], Nguyen Tien Huy[9,*]

[1] *School of Biomedical sciences, Faculty of Life Sciences, University of Bradford, Bradford, United Kingdom*

[2] *Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Tanta University, Egypt.*

[3] *Online Research Club (www.onlineresearchclub.org)*

[4] *Faculty of Medicine, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh, Viet Nam*

[5] *Faculty of Medicine, October 6 University, Giza, Egypt*

[6] *Faculty of Medicine, South Valley University, Qena, Egypt.*

[7] *Faculty of Medicine, Menoufia University, Menoufia, Egypt*

[8] *Pneumocare, Gesves, Belgium*

[9] *School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki 852-8523, Japan.*

⊕*Authors equally contributed to the work.*

Corresponding author: Nguyen Tien Huy, School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki 852-8523, Japan (E-Mail: tienhuy@nagasaki-u.ac.jp)

SM: sara.morsy@med.tanta.edu.eg **(0000-0002-2477-1139)**

THH: hhieu.truong@gmail.com (ORCID: 0000-0002-6846-8369)

AMM: abd-makram@hotmail.com (ORCID: 0000-0003-2011-8092)

OGH: osamagamal4842@gmail.com (ORCID: 0000-0002-0830-0767)

NTMD: minhuc1298@gmail.com (ORCID: 0000-0002-9333-7539)

AHZ: ahmad_zayan@yahoo.com (ORCID: 0000-0003-2581-0459)

LDNN: bacsinam81@gmail.com

NTH:tienhuy@nagasaki-u.ac.jp(ORCID:0000-0002-9543-9440)

Abstract

Purpose

Applying machine learning in medical statistics offers more accurate prediction models. In this paper, we aimed to compare the performance of the Cox Proportional Hazard model (CPH), Classification and Regression Trees (CART), and Random Survival Forest (RSF) in short-, and long-term prediction in glioblastoma patients.

Methods

We extracted glioblastoma cancer data from the Surveillance, Epidemiology, and End Results database (SEER). We used the CPH, CART, and RSF for the prediction of 1- to 10-year survival probabilities. The Brier Score for each duration was calculated, and the model with the least score was considered the most accurate.

Results

The cohort included 26473 glioblastoma patients divided into two groups: training (n = 18538) and validation set (n = 7935). The average survival duration was seven months. For the short- and long-term predictions, RSF was the best algorithm followed by CPH and CART.

Conclusion

For big data, RSF was found to have the highest accuracy and best performance. Using the accurate statistical model for survival prediction and prognostic factors determination will help the care of cancer patients. However, more developments of the R packages are needed to allow more illustrations of the effect of each covariate on the survival probability.

Keywords

Random survival forests; CART; Cox proportional hazard model; Big data; Artificial intelligence; Glioblastoma; Survival; brier score

Introduction

Malignant brain tumors are among the most formidable types of cancer, with their poor prognosis and the direct influence on cognitive functions, working ability, quality of life [1]. In 2010, according to the prevalence estimate, nearly 200000 patients were diagnosed with primary malignant brain tumors in the United States [2]. Among all the primary malignant brain tumors, malignant gliomas are the most common type with 80% of patients and an annual incidence of 5.26 per 100 000 population, which also means 17000 new cases diagnosed per year [3]. This disease, however rare in children, may present at any age, but it peaks in the sixth through eighth decades of life. Moreover, the number of patients is expected to increase with the aging of the population [3-5].

Glioblastoma (GB) is the most frequent subtype that comprises 51% of all gliomas [6]. Because of its location in the brain, aggressiveness, and low survival duration, GB is well-known as one of the most lethal cancers [7]. The common symptoms in the last month of life include seizures, headache, drowsiness, dysphagia, and eventually death rattle, agitation, and delirium. In the last stage of the disease, patients need appropriate palliative care to allow them to experience a peaceful death despite their severe symptoms [7]. According to 2016 CNS WHO classification, glioblastomas are separated into glioblastoma, IDH-wildtype (presenting in 90 % of cases), which conforms most frequently with the clinically defined primary and prevails in patients over 55 years of age [1, 8]; glioblastoma, IDH-mutant (in about 10 % of cases), which conforms closely to secondary glioblastoma and tendentiously appears in younger patients [8]; and another is glioblastoma, NOS (not otherwise specified), a diagnosis for those tumors which full IDH evaluation cannot be performed [1].

In research, especially epidemiological topics, scientists often encounter multilevel or hierarchical data, such as evaluating the potential characteristics of patients, hospitals, and regions related to the risk of death in those patients with glioblastoma during a specified duration. Survival analysis, thus, refers to methods for the analysis of data in which the outcome demonstrates the time to the onset or occurrence of a targeted event. This method of analysis has the characteristic of censorship: the event may not occur for all subjects before the completion of the study and, at the end of the study, those event-free subjects are said to be censored [9, 10].

The most basic method for survival analysis is survival tables [9, 10]. The time is divided into intervals. For each interval, the count and proportion of each living, censored and death cases are calculated. The most widely used method is the Cox Proportional Hazard regression model or approach (CPH). It estimates the magnitude of the risk of death and its confidence interval. It is used for multiple analyses of survival time data. It is considered a multiple linear regression analysis. CPH analysis depends on the assumption of the proportionality of survival time data [10]. The results are hazard ratios that estimate the probability of an event at a specific type. One of the most common cons of CPH is the convergence or the divergence of the model. This occurs if the assumption of proportionality is not fulfilled or in the presence of many covariables that are not important. It is also reported that CPH analysis had overfitting problems, which means that it describes the random error instead of examining the relationship between the variables [10].

With the development of different machine learning algorithms, new algorithms were used to deal with different limitations and problems in biomedical research. In survival analysis, many algorithms have been used. One of them is recursive partitioning algorithms. These algorithms include Random Survival Forests (RSF) and Classification and Regression Trees (CART) [11-13].

The CART algorithms have gained popularity over time due to easy implementation and interpretability. It is based upon using important variables to split the data into many nodes representing the predictors. Impurity parameters are used to identify whether to select the splitting variable and whether to continue or stop the splitting. In each node, there is a strong association between splitting variables and the response variables evidenced by the highest impurity reduction [14, 15].

Meanwhile, an RSF is an algorithm of an ensemble of trees; it is the average prediction of all trees that would produce a more accurate prediction. It is a robust algorithm against overfitting and resistant to outliers and high dimensionality data. It is a nonparametric method that can be used on any variables regardless of the distribution they follow [12, 13, 15].

In this study, we aim to compare the performance of the Cox Proportional Hazard approach (CPH), CART, and Random Survival Forest (RSF) in short-, and long-term prediction in glioblastoma patients.

Methods

Data collection

We extracted the data of cases diagnosed with glioblastoma as reported in the histology recode of brain grouping in the SEER database. We used the last version of the published US research data (1975 – 2018) released on April 2021 [16].

This included glioblastoma, NOS (9440/3), Gliosarcoma (9442/3), and giant cell glioblastoma (9441/3). SEER*stat 8.3.9.2 was used to extract the data [17]. We only included cases that died due to the tumor itself and survival time more than zero. Cases with a survival time of zero were excluded.

Based on literature, long-term survival was defined in literature as survival of glioblastoma patients more than two years [18].

Statistical analysis

The categorical variables were expressed in percentage while the continuous variables are expressed using mean and standard deviation if the data are normally distributed; otherwise, median, and interquartile range. The dataset was divided randomly into training groups with 70% of the cases and a test set with 30% of the dataset using the caret package. In the training set, missing data were imputed using the K nearest neighbors (KNN) algorithm with the number of neighbors equal to 3. The imputation was conducted in R using the VIM package [19]. For the validation set, we did sensitivity analysis where we compared the performance results between

i) Imputation of the missing data of the validation set using KNN
ii) Omitting the missing data from the validation set

Three survival algorithms were performed and compared based on an Brier Score (BS) for survival analysis on intervals of 12 months to compare the performance of the models for short- and long-term prediction. Univariable Cox Proportional Hazard regression analysis (CPH) was first developed to detect the significant covariables which were used in the multivariable analysis. Then, the accuracy of the model was assessed on the validation set using Brier score.

Random Survival Forests (RSF) were applied using the randomforestSRC package in R [20, 21]. We used 500 trees with three variables to split at each node. The variable importance

was detected using the permutation method. The predictive performance of the random forest was assessed using brier score on the validation data. The p-value for significant important variables was calculated based on Altmann et al. that depends on permutation importance [22].

A CART survival decision tree was constructed using the Rpart package; we used the minimum variable at each split of 10 and a maximum depth of 10 then we pruned the tree to avoid overfitting. Results were considered significant when the *p-value* was less than 0.05.

Brier score

In this paper, the Brier score was used to compare the accuracy of prediction of each model. The Brier score measures the accuracy of the prediction. Brier score is an evaluation metric that calculates the weighted average of squared error between the event status at time t and predicted survival probability at time t. The higher the value of the Brier score, the less accurate the results are [23, 24]. The scores were calculated using the ipred package [25]

Results

Patients' characteristics

The cohort included 26473 glioblastoma patients including training (n = 18538) and validation set (n = 7935). The median age of patients was 64 years old (Table 1). White males had the highest rates of glioblastoma. The most common site for glioblastoma was the frontal lobe. The median survival time for all patients was eight months. The cohort included patients who had anaplastic undifferentiated tumors (n = 7370). 77.4% of patients received surgical treatment; 93.4% of patients did not survive.

Survival analysis using Cox Proportional Hazard regression (CPH)

The univariable analysis revealed that all patients characteristics except gender of the patients had increased or decreased the mortality rate, for instance, older age patients had low survival probability [HR = 1.2, 95% CI (0.87, 1.5), p-value <0.001]. For tumor characteristics, pathology, different sites of the tumor, Grade, and laterality had significantly affected the survival of glioblastoma patients (Table 2). In the multivariable analysis, old white patients were diagnosed with undifferentiated tumors in the optic nerve with more than one primary tumor (Table 2). Moreover, Gliosarcoma had decreased survival probability (Table 2).

The accuracy of the model was checked using the Brier Score (BS) for both short- and long-term predictions. We found that the accuracy of the model was the least in the first year, followed by two years. After the sixth year, the model had the best accuracy that was sustained until 10-year predictions (Table 3). The same was found in the separate analysis with the validation set imputed through KNN. For machine learning algorithms.

Ensemble trees were constructed using 500 trees with three variables used at each split. The trees identified age, laterality, and surgery as the top three important variables for the prediction of patient survival (Figure 1). Random forest survival trees had high accuracy (low brier score); the BS was 0.177 for the first year and decreased for the long-term analysis (0.01 for 10 years' prediction) Table 3.

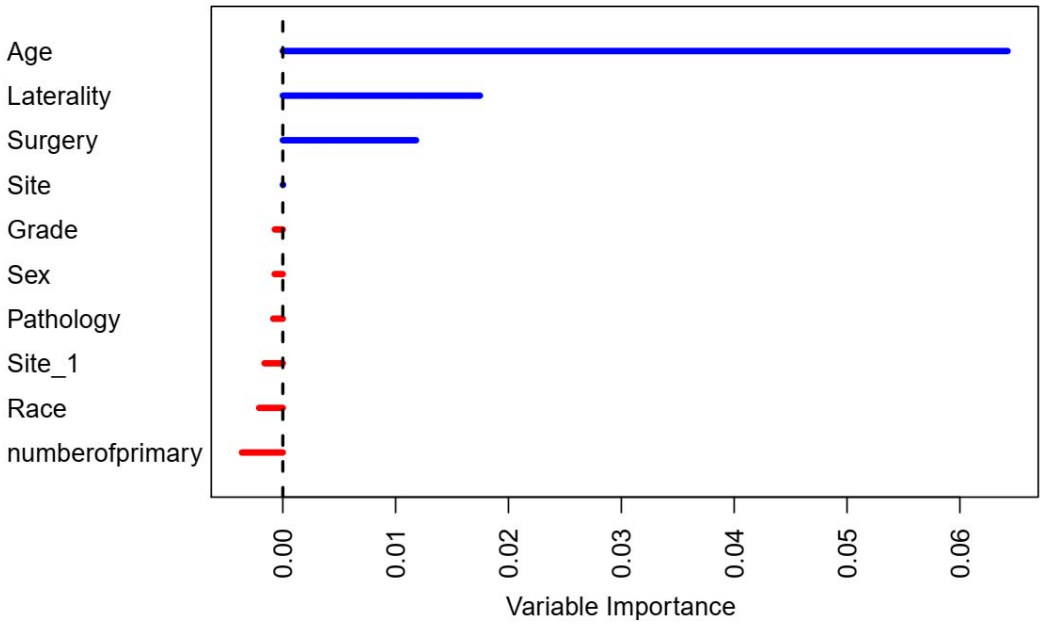

Figure 1. Variable importance plot illustrating the top variables used for prediction. Site_1 is the specific central nervous system sites.

For the CART survival tress that used all the predictors for survival prediction, the cohort was divided into six groups as shown in (Figure 2). The groups are: 1) Patients aged more

than 66.5 and performed surgery (median survival = 6 months, n = 5079, 2) Patients aged more than 66.5 and did not perform any surgery (median survival = 3 months, n = 1872), 3) Patients who aged less than 45.5 (median survival = 16 months, n = 2283) have survival probability of 0.538, 4) Patients whose age are between 45.5 – 66.5 and performed surgery (median survival = 11 months, n = 7271), 5) Patients whose age are between 45.5 – 66.5 and did not perform surgery (median survival = 5 months, n = 1578)

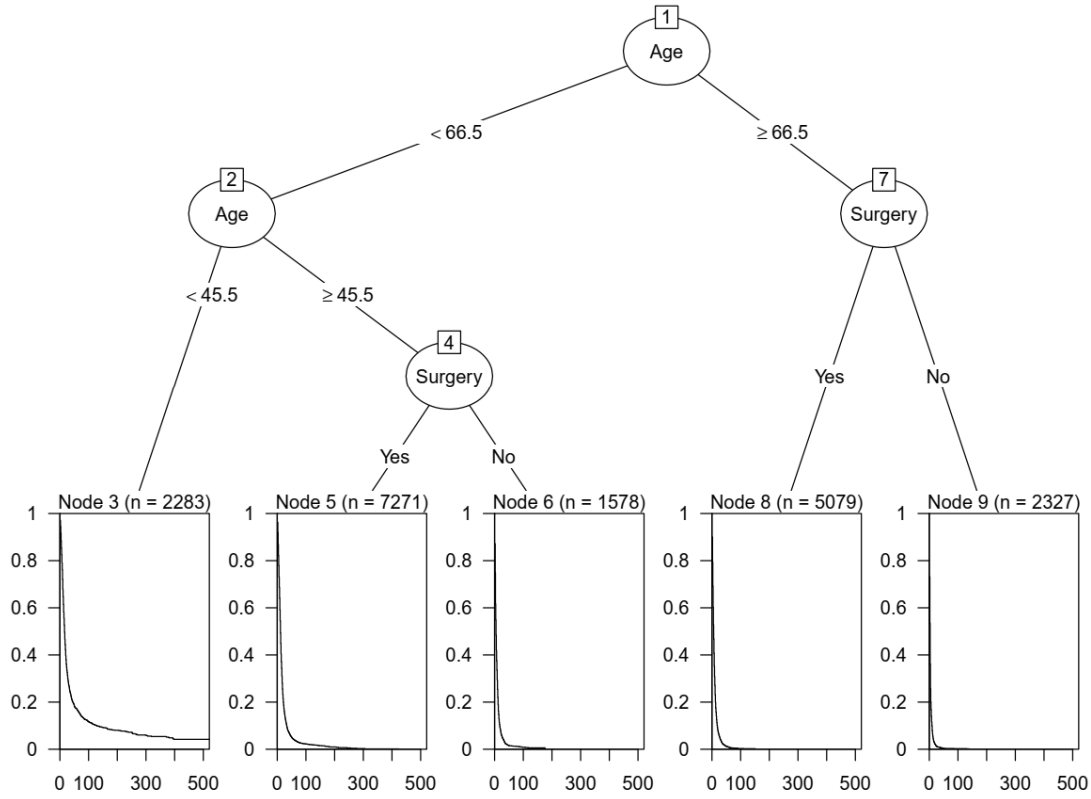

Figure 2. The recursive partitioning survival trees illustrating the survival probability for each group of patients.

Assessing the accuracy of the CART model using BS revealed that CART has much less accuracy than CPH and random forest implying that it is not suitable for long-term prediction Table 3. The accuracy of the model is decreasing through time with the highest accuracy in the first year (Figure 3).

Based on our results, we found that random forest maintained high accuracy (low Brier score) for both short- and long-term predictions followed by CPH followed by CART survival trees Figure 3. Brier score for CPH model decreased for longer duration indicating the high

accuracy of model for long term prediction, however, Brier score was higher than random forest suggesting random forest is the most suitable model for short- and long-term predictions Table 3. Recursive partitioning trees showed very low accuracy compared to other models.

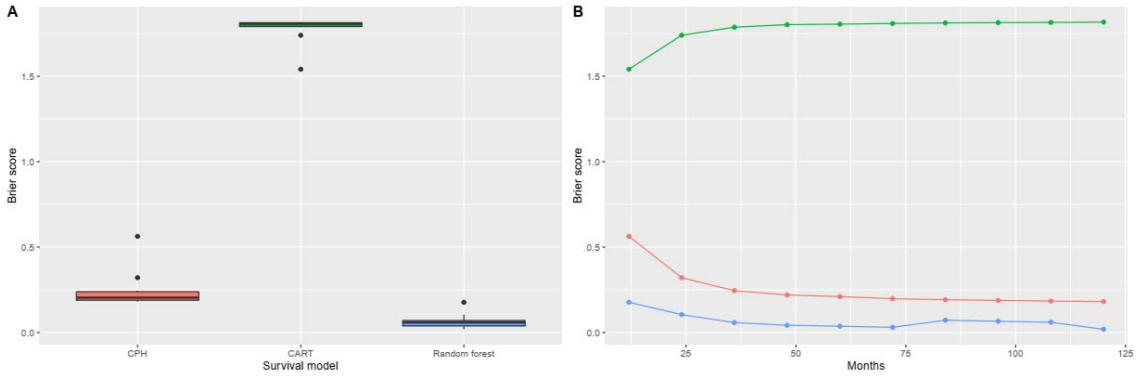

Figure 3. Boxplot showing the overall median brier score for each model suggesting random forest had the highest accuracy (A), Figure B shows the brier score at each time point indicating that random forest had the lowest brier score hence the highest accuracy

Discussion

This study aimed at investigating the appropriate survival models to use for the big data in medicine. We compared three models using Brier Scores (BS) as an indicator of model accuracy. Based on our results, RSF had the best accuracy for short- and long-term predictions followed by CPH.

Machine learning introduction into medical sciences and data analytics has created a massive impact in the field of public health [26]. With the help of automated processes and artificial intelligence, we acquired the ability to read unrevealed data and algorithm patterns.

According to Nasejje et al., when faced with irrelevance between the covariates in the model for time-to-event data, they use the Cox model and link each covariate with one public health assumption [27]. On the other hand, when using RSF, the data will be not reliant on the assumption for its validity [28]. So, Nasejje et al. used the RSF to analyze public health data to figure out associated factors with mortality of younger group of patients (under the age of five) [27]. Typically, in any dataset that includes many covariables carrying the risk of violating the proportional hazards assumption, RSF can be considered [29, 30].

Moreover, Random Survival Forests (RSF), in several risk models, outperformed the traditional Cox Proportional Hazard model (CPH). More importantly, while Cox cannot automatically identify the nonlinear effects of all considered variables, RSF can [28, 30-33]. However, in 2009, on breast cancer patients, the authors reported that CPH was a reliable method for predicting disease-free survival (DFS) in cancer. It was more advantageous than RSF approaches. This was justified by the capability of CPH to extract patterns and relationships hidden deep into medical datasets, leading to high predictive abilities that can be used for different sample sizes and potential future suitable survival data problems, whereas RSF provides only interpretive results [34].

For a dataset with separated and different risk levels, “Model-Based Recursive Partitioning” was able to present a good description. Safe M et al. reported the superiority of recursive partitioning for nonlinear model structures [35]. In the interaction dataset, recursive partitioning like Classification and Regression Trees (CART) and Artificial Neural Network (ANN) showed superior results to Cox ($P< 0.05$) with an improvement of 0.1 (95% CI, 0.08 to 0.12) and 0.015 (95% CI, 0.01 to 0.02) respectively. In theory, CART and ANN overcame the limitations of the Cox model regarding the ease and extent of their use [36]. In a study about breast tumor chemosensitivity to primary chemotherapy, on the other hand, the logistic

regression model predictions were better than recursive partitioning [37, 38]. Another study by Lee et al. confirmed that Cox linear regression modeling outperformed recursive partitioning when there were only continuous predictors, while recursive partitioning was better when there were significant categorical predictors [37, 39]. One last study by Chun et al. demonstrated that Artificial Neural Network (ANN) had worse performance than the logistic regression model [37, 40]. Because the main goal of those methods (RSF, CART, and ANN) is to develop a predictive model of many variables that can lead to a more efficient clinical use that is particularly important for physicians. Contrastingly, conventional statistical modeling needs proper input of data from an expert to create a much easier model to interpret than data-driven-based techniques. This, in turn, leads to the narrow scope of conventional techniques to find new correlations between the data ready to be used in the literature [41].

Another study addressing dyslipidemia analyzed the difference in the disease incidence using RSF and Cox model and summarized that the RSF could predict more variables than the CPH. Those variables included the baseline lipoprotein profiles (including high- and low-density lipoproteins, total cholesterol, total triglycerides, blood pressure, age, body mass index, … etc.) [42]. Accordingly, not only do we need a tool to analyze the mortality, morbidity, and risk rates, but also a tool we can depend on searching for more variables effectively. So, using machine learning techniques can help us achieve combinations that we can barely capture using conventional approaches [41].

In our models, we validated our results using Brier Score (BS). An earlier study published in 2009 compared the RSF to CPH, using the Harrell c-statistics [43-45] for assessing the validity, found that CPH had better performance, and concluded that its replacement by RSF is still controversial and needed further investigation [46]. On the other hand, two other studies, assessing the 1-year mortality and survival in cardiac patients with cardiac arrhythmias, compared the use of RSF and the traditional CPH. The results were that RSF significantly overperformed CPH [47, 48]. The latter findings were supported by numerous other published studies [31, 42, 49, 50].

One of the biggest obstacles is not only to collect datasets that have the appropriate size and needed quality of samples but also to use appropriate methods in analysis. Towards that point, our study contributes as it compares these predictive methods which in turn can improve and reinforce the theory about the limitations and advantages of each method [26].

Limitation

Our analyses were conducted in R. The effect of each level of the categorical variables in the survival trees could not be aptly illustrated due to the lack of R packages with more advanced illustrations. Another limitation is that random forest and the function used to calculate brier score failed if the data has missing values.

Conclusion

In this paper, we compared the performance of the Cox Proportional Hazard model (CPH), Classification and Regression Trees (CART), and the Random Survival Forest (RSF) in predicting the survival of glioblastoma patients reported in the SEER database. We concluded that the RSF achieved the best performance and the highest accuracy followed by the CPH and lastly by the CART for both short- and long-term predictions, validated by the Brier Score (BS). Accordingly, using RSF may be of benefit in determining the best prognostic factors of cancer patients; however, more development of R packages is needed to allow for more illustrations of each covariate effect. More studies of the same kind are also needed to examine the performance of the three models in other cancer types.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

The authors of this manuscript consent to the publication of the work by BioMed Central.

Availability of data and material

The data are available and can be accessed through the SEER database, which is publicly available at https://seer.cancer.gov/data/.

Competing interests

None

Funding

None

Authors' Contributions

SM is responsible for the idea of the study, concept and design, acquisition of the data from the SEER database, statistical analysis and interpretation of data. SM, AMM, HT, OGH, NTMD, LDNN, and AHZ contributed to writing the manuscript. The study was under the supervision of NTH, who has also revised the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

None

Code availability

The codes used in the analysis are available upon request from Dr. Sara Morsy (sara.morsy@med.tanta.edu.eg)

References

1. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta neuropathologica. 2016;131(6):803-20.
2. Porter KR, McCarthy BJ, Freels S, Kim Y, Davis FG. Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology. Neuro-oncology. 2010;12(6):520-7.
3. Dolecek TA, Propp JM, Stroup NE, Kruchko C. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005-2009. Neuro-oncology. 2012;14 Suppl 5:v1-49.
4. Wohrer A, Waldhor T, Heinzl H, Hackl M, Feichtinger J, Gruber-Mosenbacher U, et al. The Austrian Brain Tumour Registry: a cooperative way to establish a population-based brain tumour registry. Journal of neuro-oncology. 2009;95(3):401-11.
5. Ostrom QT, Gittleman H, Stetson L, Virk SM, Barnholtz-Sloan JS. Epidemiology of gliomas. Cancer treatment and research. 2015;163:1-14.
6. Kleihues P, Burger PC, Scheithauer BW. The new WHO classification of brain tumours. Brain pathology (Zurich, Switzerland). 1993;3(3):255-68.
7. Preusser M, de Ribaupierre S, Wohrer A, Erridge SC, Hegi M, Weller M, et al. Current concepts and management of glioblastoma. Annals of neurology. 2011;70(1):9-21.
8. Ohgaki H, Kleihues P. The definition of primary and secondary glioblastoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(4):764-72.
9. Austin PC. A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications. International statistical review = Revue internationale de statistique. 2017;85(2):185-203.
10. George B, Seals S, Aban I. Survival analysis and regression models. Journal of nuclear cardiology : official publication of the American Society of Nuclear Cardiology. 2014;21(4):686-94.
11. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Statistics Surveys. 2011;5:44-71.
12. Ishwaran H, Lu M. Random survival forests. Wiley StatsRef: Statistics Reference Online. 2007:1-13.

13. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. Biostatistics. 2005;7(3):355-73.

14. Barlin JN, Zhou Q, St Clair CM, Iasonos A, Soslow RA, Alektiar KM, et al. Classification and regression tree (CART) analysis of endometrial carcinoma: Seeing the forest for the trees. Gynecol Oncol. 2013;130(3):452-6.

15. Zhou Y, McArdle JJ. Rationale and Applications of Survival Tree and Survival Ensemble Methods. Psychometrika. 2015;80(3):811-33.

16. National Cancer Institute D, Surveillance Research Program. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties. released April 2021, based on the November 2020 submission.

17. Surveillance Research Program, National Cancer Institute SEER*Stat software (seer.cancer.gov/seerstat) version 8.3.9.2.

18. Poon MTC, Sudlow CLM, Figueroa JD, Brennan PM. Longer-term (≥□2 years) survival in patients with glioblastoma in population-based studies pre- and post-2005: a systematic review and meta-analysis. Scientific Reports. 2020;10(1):11622.

19. Kowarik A, Templ M. Imputation with the R Package VIM. Journal of statistical software. 2016;74(7):1 - 16.

20. Ishwaran H, Kogalur UB. Random survival forests for R. R news. 2007;7(2):25-31.

21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2(3):841-60.

22. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340-7.

23. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biometrical journal Biometrische Zeitschrift. 2008;50(4):457-79.

24. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999;18(17-18):2529-45.

25. Hothorn APaT. ipred: Improved Predictors. R package version 0.9-9. 2019.

26. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.

27. Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. BMC Res Notes. 2017;10(1):459.

399 28. Nasejje JB, Mwambi HG, Achia TN. Understanding the determinants of under-five
400 child mortality in Uganda including the estimation of unobserved household and community
401 effects using both frequentist and Bayesian survival analysis approaches. BMC Public
402 Health. 2015;15:1003.
403 29. Ehrlinger J. ggRandomForests: Exploring Random Forest Survival. arXiv. 2016.
404 30. Taylor JM. Random Survival Forests. J Thorac Oncol. 2011;6(12):1974-5.
405 31. Datema FR, Moya A, Krause P, Back T, Willmes L, Langeveld T, et al. Novel head
406 and neck cancer survival analysis approach: random survival forests versus Cox proportional
407 hazards regression. Head Neck. 2012;34(1):50-8.
408 32. Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying Important Risk Factors for
409 Survival in Kidney Graft Failure Patients Using Random Survival Forests. Iranian journal of
410 public health. 2016;45(1):27-33.
411 33. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important
412 risk factors for survival in patient with systolic heart failure using random survival forests.
413 Circulation Cardiovascular quality and outcomes. 2011;4(1):39-45.
414 34. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT, et al.
415 Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal
416 women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a
417 retrospective analysis of a randomised trial. The Lancet Oncology. 2010;11(1):55-65.
418 35. Safe M, Faradmal J, Mahjub H. A Comparison between Cure Model and Recursive
419 Partitioning: A Retrospective Cohort Study of Iranian Female with Breast Cancer. Comput
420 Math Methods Med. 2016;2016:9425629-.
421 36. Kattan MW, Hess KR, Beck JR. Experiments to Determine Whether Recursive
422 Partitioning (CART) or an Artificial Neural Network Overcomes Theoretical Limitations of
423 Cox Proportional Hazards Regression. Computers and Biomedical Research. 1998;31(5):363-
424 73.
425 37. Ballester M, Oppenheimer A, Mathieu d'Argent E, Touboul C, Antoine J-M, Coutant
426 C, et al. Nomogram to predict pregnancy rate after ICSI–IVF cycle in patients with
427 endometriosis. Human reproduction. 2011;27(2):451-6.
428 38. Rouzier R, Coutant C, Lesieur B, Mazouni C, Incitti R, Natowicz R, et al. Direct
429 comparison of logistic regression and recursive partitioning to predict chemotherapy response
430 of breast cancer based on clinical pathological variables. Breast cancer research and
431 treatment. 2009;117(2):325-31.

39. Lee JW, Um SH, Lee JB, Mun J, Cho H. Scoring and Staging Systems Using Cox Linear Regression Modeling and Recursive Partitioning. Methods Inf Med. 2006;45(01):37-43.

40. Chun FKH, Graefen M, Briganti A, Gallina A, Hopp J, Kattan MW, et al. Initial Biopsy Outcome Prediction—Head-to-Head Comparison of a Logistic Regression-Based Nomogram versus Artificial Neural Network. European Urology. 2007;51(5):1236-43.

41. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLOS ONE. 2018;13(8):e0202344.

42. Zhang X, Tang F, Ji J, Han W, Lu P. Risk Prediction of Dyslipidemia for Chinese Han Adults Using Random Forest Survival Model. Clin Epidemiol. 2019;11:1047-55.

43. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.

44. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984;3(2):143-52.

45. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361-87.

46. Kurt Omurlu I, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. Expert Systems with Applications. 2009;36(4):8582-8.

47. Miao F, Cai Y-P, Zhang Y-X, Li Y, Zhang Y-T. Risk Prediction of One-Year Mortality in Patients with Cardiac Arrhythmias Using Random Survival Forest. Comput Math Methods Med. 2015;2015:303250-.

48. Miao F, Cai Y-P, Zhang Y-T, Li C-Y, editors. Is Random Survival Forest an Alternative to Cox Proportional Model on Predicting Cardiovascular Disease? 6th European Conference of the International Federation for Medical and Biological Engineering; 2015 2015//; Cham: Springer International Publishing.

49. Dietrich S, Floegel A, Troll M, Kuhn T, Rathmann W, Peters A, et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol. 2016;45(5):1406-20.

50. Roshanaei G, Omidi T, Faradmal J, Safari M, Poorolajal J. Determining affected factors on survival of kidney transplant in living donor patients using a random survival forest. Koomesh. 2018;20(3):517-23.

466

Table 1. Characteristics of the included patients

	level	Overall	Training	Validation
n		26473	18538	7935
Age (median [IQR])		63.00 [54.00, 72.00]	63.00 [54.00, 72.00]	83.54 (133.34)
Sex (%)	Female	11161 (42.2)	7851 (42.4)	3310 (41.7)
	Male	15312 (57.8)	10687 (57.6)	4625 (58.3)
Race (%)	Black	1260 (4.8)	904 (4.9)	356 (4.5)
	Other (American Indian/AK Native, Asian/Pacific Islander)	1144 (4.3)	803 (4.3)	341 (4.3)
	White	24032 (90.9)	16806 (90.8)	7226 (91.2)
Site (%)	Brain	26409 (99.8)	18491 (99.7)	7918 (99.8)
	Cranial Nerves Other Nervous System	64 (0.2)	47 (0.3)	17 (0.2)
Specific Site (%)	Brain stem	166 (0.6)	117 (0.6)	49 (0.6)
	Brain, NOS	1908 (7.2)	1336 (7.2)	572 (7.2)
	Cauda equina	2 (0.0)	2 (0.0)	51 (0.6)
	Cerebellum, NOS	173 (0.7)	122 (0.7)	309 (3.9)
	Cerebrum	1031 (3.9)	722 (3.9)	1 (0.0)
	Cranial nerve, NOS	4 (0.0)	3 (0.0)	2003 (25.2)
	Frontal lobe	6677 (25.2)	4674 (25.2)	333 (4.2)
	Occipital lobe	1113 (4.2)	780 (4.2)	1 (0.0)
	Optic nerve	6 (0.0)	5 (0.0)	1339 (16.9)
	Overlapping lesion of brain	4466 (16.9)	3127 (16.9)	1340 (16.9)
	Parietal lobe	4467 (16.9)	3127 (16.9)	15 (0.2)
	Spinal cord	52 (0.2)	37 (0.2)	1891 (23.8)
	Temporal lobe	6304 (23.8)	4413 (23.8)	31 (0.4)
	Ventricle, NOS	104 (0.4)	73 (0.4)	18 (0.7)
Grade (%)	Moderately differentiated; Grade II	57 (0.7)	39 (0.7)	276 (11.1)
	Poorly differentiated; Grade III	871 (10.5)	595 (10.2)	2181 (88.0)

	Undifferentiated; anaplastic; Grade IV		7370 (88.6)	5189 (88.8)	4 (0.2)
	Well differentiated; Grade I		22 (0.3)	18 (0.3)	1600 (20.2)
Laterality (%)	Left		5280 (20.0)	3680 (19.9)	2 (0.0)
	midline tumor		17 (0.1)	15 (0.1)	4580 (57.9)
	Not a paired site		15261 (57.8)	10681 (57.8)	1671 (21.1)
	Right		5641 (21.4)	3970 (21.5)	62 (0.8)
	Single not specified		191 (0.7)	129 (0.7)	1800 (23.2)
Surgery (%)	No		5850 (22.6)	4050 (22.3)	5963 (76.8)
	Yes		20037 (77.4)	14074 (77.7)	14.06 (27.74)
Survival months (median [IQR])			8.00 [3.00, 15.00]	8.00 [3.00, 15.00]	531 (6.7)
status (%)		Dead	1750 (6.6)	1219 (6.6)	7404 (93.3)
		Alive	24723 (93.4)	17319 (93.4)	1.16 (0.45)
Number of primary (median [IQR])			1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Pathology (%)	Giant cell glioblastoma		259 (1.0)	197 (1.1)	62 (0.8)
	Glioblastoma, NOS		25700 (97.1)	17995 (97.1)	7705 (97.1)
	Gliosarcoma		514 (1.9)	346 (1.9)	166 (2.1)

IQR: interquartile range

467

Table 2. Univariable and multivariable cox proportional hazard model

Characteristic	N	HR[1]	95% CI[1]	p-value	HR[1]	95% CI[1]	p-value
Age	18,538	1.2	0.87, 1.5	<0.001	1	1.00, 1.00	<0.001
Sex	18,538						
Female		—	—				
Male		0.98	0.95, 1.01	0.17			
Race	18,538						
Black		—	—		—	—	
Other (American Indian/AK Native, Asian/Pacific Islander)		1.04	0.94, 1.15	0.47	1.05	0.95, 1.16	0.4
White		1.24	1.15, 1.33	<0.001	1.22	1.13, 1.30	<0.001
Site	18,538						
Brain		—	—		—	—	
Cranial Nerves Other Nervous System		0.71	0.52, 0.97	0.031			
Specific Sites	18,538						
Brain stem		—	—		—	—	
Brain, NOS		1.58	1.30, 1.93	<0.001	1.75	1.44, 2.14	<0.001
Cauda equina		0.25	0.03, 1.79	0.17	0.35	0.05, 2.48	0.3
Cerebellum, NOS		0.98	0.75, 1.28	0.87	1.31	1.00, 1.71	0.049
Cerebrum		1.36	1.11, 1.67	0.003	1.68	1.37, 2.06	<0.001
Cranial nerve, NOS		3.74	1.19, 11.8	0.024	3.33	1.06, 10.5	0.04
Frontal lobe		0.98	0.81, 1.19	0.84	1.53	1.26, 1.86	<0.001
Occipital lobe		1.01	0.82, 1.24	0.92	1.56	1.27, 1.91	<0.001
Optic nerve		2.07	0.84, 5.07	0.11	1.34	0.54, 3.29	0.5
Overlapping lesion of brain		1.29	1.06, 1.56	0.01	1.64	1.35, 1.99	<0.001
Parietal lobe		1.07	0.88, 1.30	0.5	1.64	1.35, 1.99	<0.001
Spinal cord		0.7	0.47, 1.04	0.08	0.88	0.59, 1.31	0.5

Temporal lobe		1.01	0.83, 1.22	0.94	1.59	1.31, 1.93	<0.001
Ventricle, NOS		0.97	0.71, 1.31	0.83	1.3	0.96, 1.77	0.091
Grade	18,538						
Moderately differentiated; Grade II		—	—		—	—	
Poorly differentiated; Grade III		1.39	1.03, 1.88	0.03	1.4	1.04, 1.88	0.029
Undifferentiated; anaplastic; Grade IV		1.32	0.99, 1.77	0.061	1.44	1.08, 1.94	0.014
Well differentiated; Grade I		1.2	0.74, 1.93	0.46	1.42	0.88, 2.29	0.15
Laterality	18,538						
Left		—	—		—	—	
midline tumor		1.14	0.65, 2.01	0.65	1.17	0.66, 2.07	0.6
Not a paired site		1.48	1.42, 1.54	<0.001	1.45	1.39, 1.52	<0.001
Right		0.98	0.94, 1.03	0.53	1.02	0.97, 1.07	0.4
Single not specified		1.35	1.12, 1.62	0.001	1.2	0.99, 1.44	0.057
Surgery	18,538						
No		—	—		—	—	
Yes		0.47	0.45, 0.49	<0.001	0.52	0.50, 0.54	<0.001
numberofprimary	18,538	1.09	1.06, 1.13	<0.001	1.08	1.05, 1.12	<0.001
Pathology	18,538						
Giant cell glioblastoma		—	—		—	—	
Glioblastoma, NOS		1.42	1.22, 1.65	<0.001	1.27	1.09, 1.48	0.002
Gliosarcoma		1.36	1.13, 1.64	0.001	1.44	1.20, 1.74	<0.001

[1] **HR = Hazard Ratio, CI = Confidence Interval**

473

Table 3. Comparison between each model using brier score to estimate the accuracy of each model for short- and long-term prediction in test set with or without missing data imputation

Time	**Cox proportional Hazard ratio**		**CART**		**Random forest**	
	No imputation in the validation set	**KNN imputation in the validation set**	**No imputation in the validation set**	**KNN Imputation in the validation set**	**No imputation in the validation set**	**KNN imputation in the validation set**
12 months	0.563	0.658	1.541	1.609	0.177	0.173
24 months	0.321	0.392	1.740	1.801	0.105	0.152
36 months	0.245	0.309	1.787	1.851	0.058	0.126
48 months	0.220	0.279	1.802	1.866	0.0423	0.107
60 months	0.210	0.264	1.805	1.847	0.037	0.093
72 months	0.198	0.252	1.809	1.873	0.0303	0.08
84 months	0.192	0.243	1.812	1.883	0.072	0.075
96 months	0.188	0.239	1.814	1.886	0.066	0.069
108 months	0.184	0.230	1.815	1.889	0.060	0.064
120 months	0.181	0.226	1.817	1.889	0.019	0.059

474

475

Table legends

Table 1. The characteristics of included patients.

Table 2. Univariable and multivariable Cox regression analysis

Table 3. Comparison between each model using brier score for each model to estimate the accuracy of each model for short- and long-term prediction

Figure legends

Figure 1. Variable importance plot illustrating the top variables used for prediction.

Figure 2. The recursive partitioning survival trees illustrating the survival probability for each group of patients.

Figure 3. Boxplot showing the overall median brier score for each model suggesting random forest had the highest accuracy (A), Figure B shows the brier score at each time point indicating that random forest had the lowest brier score hence the highest accurracy