Analyzing Glioblastoma Patient Survival with Cox Regression

Philip Lee, Danah Albaaj, Grace Ramey, April Zhou

Abstract

Machine learning algorithms have become progressively more useful for analyzing patient data and making predictions regarding patient survival. In a recent paper, medical imaging researchers at UCLA created and tested logistic regression models that can make 2-, 6-, or 9-month residual survival predictions in glioblastoma patients.[1] Our study aims to generate a continuous residual survival prediction, instead of the discretized output of the original paper, in order to give a more precise estimate of survival time. Kaplan-Meier survival curve analysis was used to identify potentially significant covariates from a glioblastoma data set provided by The Cancer Genome Atlas (TCGA). These covariates were then inputted into a Cox regression model, which returned three significant covariates: radiation therapy type, post-operative drug treatment status, and drug duration. Based on this model, undergoing external beam radiation therapy, utilizing pharmaceutical treatment as a secondary round of treatment, and increasing drug duration should lengthen a patient's residual survival. To verify the viability of this model, 1000 simulations were run for each patient in a testing data set to get an average, "most probable" survival time for each test patient. Comparing predicted and actual survival times of test patients, we found that our model ultimately underestimated patient survival times. Still, the significant covariates identified by the Cox model should be further evaluated. Additionally, future Cox models should consider incorporating covariates identified with more sophisticated pattern-mining algorithms, such as cSPADE.

Introduction

Glioblastoma multiforme (GBM) is a fast-growing, malignant type of cancer that begins in the brain or spinal cord. It commonly recurs, with prognosis being difficult to predict.[2] Several studies have tried to improve patient prognosis prediction, using machine learning algorithms to glean patterns from patients' demographic and longitudinal data.[3,4] In a recent paper, medical imaging researchers at UCLA created logistic regression models that can predict 2-, 6-, or 9-month residual survival—the expected time from last visit or treatment to time of death—of GBM patients, with model inputs including disease trajectory and other covariates such as sex, race/ethnicity, and interventions.[1] While such prediction is valuable towards improving patient prognosis, the binary outcome of such a model (for instance, that a patient will either survive to two months or not) is somewhat limited. Thus, for our project, we sought to create a model that can make residual survival predictions for GBM patients on a continuous scale, with units of days. This is because a continuous residual survival time output would be far more useful for patients, as it is a more precise measure of the length of time they are likely to survive. Several types of survival curve analyses were used to generate a model that could give us this result; the Cox model was our ultimate focus.

Methods

Data and Initial Processing

The dataset we used (TCGA-GBM) is from a large study with an initial cohort of 262 patients and contains patients' clinical, genetic, imaging, and pathological information.[5] Covariates and patients were removed if they did not meet the appropriate minimum missing information thresholds of 90% for columns and 75% for rows. This produced a cohort of 149 patients for the training dataset for our model and a cohort of 36 patients for our testing set.

Kaplan-Meier Survival Curves

In studies predicting residual patient survival, the main dependent variable of interest is residual survival. However, this value is only available from patients who have actually died, and it is not immediately obvious how to incorporate data from patients who have not died or have been lost to follow-up. The Kaplan-Meier (K-M) method deals with data sets that have these types of "incomplete observations."[6] In this method, patients who do not experience the event (i.e., death) are labeled as "censored." The probability of survival over time is calculated for each interval of time, and censored patients who have not died or are lost to follow-up can still be taken into account in the construction of the survival curve.

In our project, we used the K-M method to identify covariates that may significantly affect patient survival. This was done by splitting the patients in our dataset into two different cohorts based on a particular covariate, plotting the survival curve for each cohort, and then using the log-rank test to evaluate whether the two curves are significantly different from each other. All K-M calculations and plots were done using the survival, survminer, and ggplot2 packages in R.

Cox Proportional Hazards Model

The Cox proportional hazards model is a multivariate regression model that allows us to predict survival time based on several covariates (categorical and/or continuous) at once.[7] The "hazard" can be interpreted as the risk of an individual dying at time t. The hazard function, $h(t)$, is the measure of risk that a patient will die at time t given that the patient has survived to time t, and given the patient's unique set of covariates. The hazard function is the following:

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) \tag{1}$$

where t represents survival time, $(x_1, x_2, \ldots, x_n)$ represent the covariate values, and $h_0(t)$ represents baseline hazard (i.e., risk of death if the values of all covariates are zero). The coefficients $(\beta_1, \beta_2, \ldots, \beta_n)$ represent the weights of each covariate. The values of β are calculated by fitting the hazard function to a survival curve for the population. Also, e^{β_i} is the hazard ratio for the i^{th} covariate, which can be interpreted as the factor by which a patient's risk is multiplied given a one-unit increase in the value of the i^{th} covariate. When e^{β_i} is greater than one, it indicates that increasing the i^{th} covariate increases the probability of death (bad prognostic factor), and when e^{β_i} is less than one, it indicates that increasing the i^{th} covariate decreases the probability of death (good prognostic factor).

Using the K-M method and univariate Cox regression (Cox proportional hazards model but with only one predictor variable), we identified several variables that appeared to affect probability of patient survival. Then, we used those variables to build a multivariate Cox regression model (using our training data). All Cox regression calculations and plots were done using the survival, survminer, and ggplot2 packages in R.

Predicting Survival Times

Applying our Cox model to each patient in our testing data set, we generated a specific hazard function for each patient, which was then converted to a specific survival function for each patient. Next, the survival function was converted to a specific cumulative distribution function (CDF). Finally, for each patient-specific CDF, we sampled 1000 survival times and calculated the average survival time and standard deviation. We evaluated the accuracy of our predicted times by performing a linear regression of predicted survival times against actual survival times of our test patients.

Results

Identifying Covariates of Interest

After the initial data processing, we were left with 35 patient covariates to choose from. We used both the K-M method and univariate Cox regression to determine covariates that appeared to significantly affect patient survival ($p < 0.05$ for their respective statistical significance tests). We narrowed down the 35 covariates to just six potentially significant covariates.

Training the Multivariate Cox Model

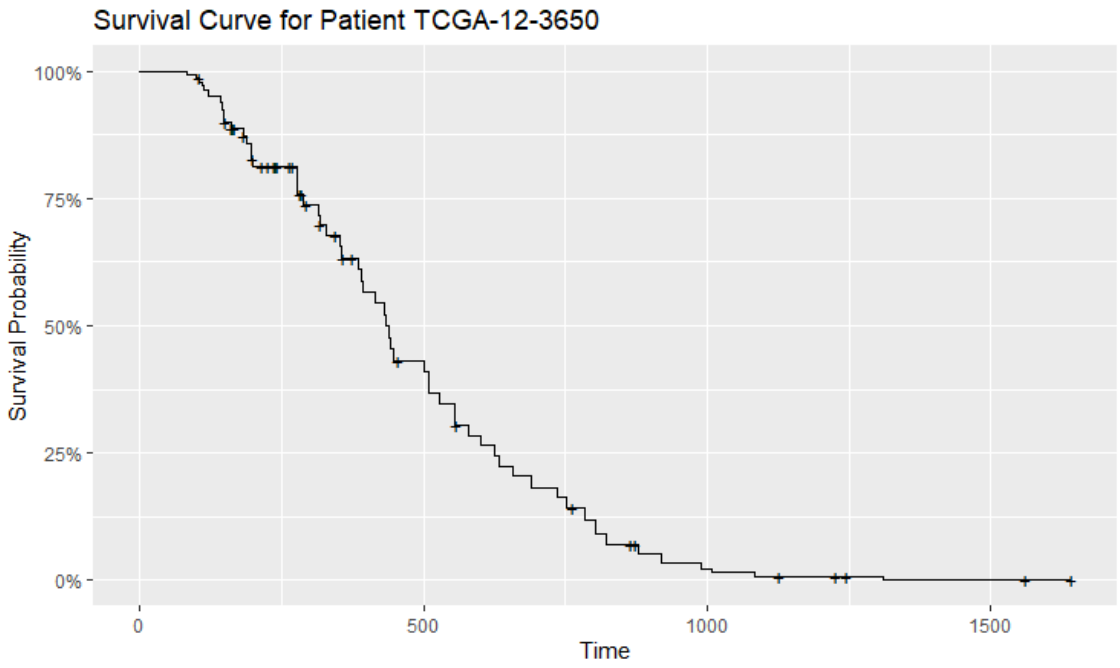
After inputting our training data into the Cox function (from the survival package in R), the six covariates that we initially marked as significant were further limited to just three statistically significant covariates:

Covariate Name	β **Coefficient**	e^{β} **Hazard Ratio**
radiation_therapy	-1.624	0.197
postoperative_rx_tx	-2.360	0.094
drug_duration	-0.006	0.994

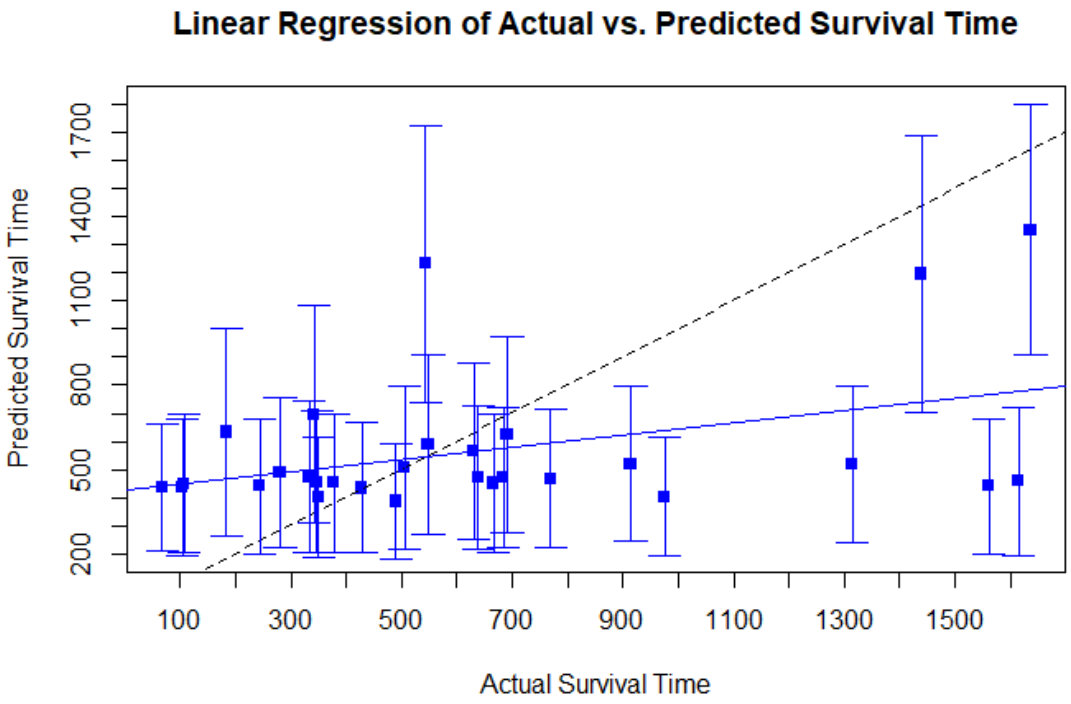
We interpreted the hazard ratio (e^{β}) outputs as follows: undergoing external beam radiation therapy will reduce risk of patient death by around 80.3%, utilizing pharmaceutical treatment as a secondary round of treatment will reduce risk of patient death by around 90.6%, and increasing drug duration by one day will reduce risk of patient death by around 0.6%. These percentages were calculated using the formula $(1 - e^{\beta}) \times 100\%$.

Testing the Multivariate Cox Model

Using our Cox model, we generated patient specific survival curves for each patient in our testing set. The following is an example of a survival curve calculated for one of our test patients:



where time on the x-axis is in days. Next, converting each patient specific survival curve into a CDF, we calculated average survival times for each patient (based on 1000 random samples from each patient's CDF). Using these average survival times as the best possible prediction for patient survival times, the linear regression of predicted times against actual times for our testing patients was the following:



The best-fit line of an ideal model should have a slope of 1, with an R^2-value also close to 1. The black dotted line in the plot is the ideal best-fit line. The blue line represents the best-fit line for our model, which has a slope of around 0.216 and an R^2-value of 0.159. The error bars represent one standard deviation above and below the average predicted survival time for each patient's 1000 simulations. Since the slope of the best-fit line was less than 1, overall, the model tended to underestimate survival times (especially for patients who had long residual survivals). However, many of the survival intervals of our test patients did overlap with the ideal line.

Discussion and Conclusion

The Cox proportional hazards regression model in this project was our attempt to generate a continuous patient survival time predictions using multiple patient covariates. However, the model produced from the data set we had available to us could not accurately predict patient residual survival (tendency to underestimate survival times). There are a few reasons why this may be the case: first, the data was rather limited. Only 90 patients out of the anticipated 149 patients had information available for training and testing the model with the covariates of interest. Second, our covariate filtering methods only gave us three covariates to input into the Cox model. The data set we had did not include certain covariates that the original paper labeled as significant, such as patient tumor volume over time. Finally, generating an accurate continuous prediction for such a complex condition is extremely difficult when working with a limited data set; even the authors who had access to data of their choosing in the original study we based our project off of opted for a model that generated discrete survival time predictions.

Despite the poor performance of the multivariate Cox model, our results from the K-M and univariate Cox regression analyses suggest the importance of covariates like drug duration and radiation therapy in determining patient survival. These covariates should be investigated further and used when constructing survival models in the future. Temporal patterns identified with more sophisticated algorithms, such as the pattern-mining algorithm cSPADE, should be incorporated as well to give the most accurate survival prediction possible for patients with GBM.

Proof of 3-Page Limit

https://docs.google.com/document/d/13Bov4CY7KkyjQ5Bahc2OlgDaPfcGSRIkYXuT1T6yEGo/edit?usp=sharing

References

1. Smedley, N. F., Ellingson, B. M., Cloughesy, T. F., & Hsu, W. (2018). Longitudinal Patterns in Clinical and Imaging Measurements Predict Residual Survival in Glioblastoma Patients. Scientific Reports, 8(1), 14429. https://doi.org/10.1038/s41598-018-32397-z

2. Omuro, A., & DeAngelis, L. M. (2013). Glioblastoma and other malignant gliomas: A clinical review. JAMA, 310(17), 1842–1850. https://doi.org/10.1001/jama.2013.280319

3. Barthel, F. P., Johnson, K. C., Varn, F. S., Moskalik, A. D., Tanner, G., Kocakavuk, E., Anderson, K. J., Abiola, O., Aldape, K., Alfaro, K. D., Alpar, D., Amin, S. B., Ashley, D. M., Bandopadhayay, P., Barnholtz-Sloan, J. S., Beroukhim, R., Bock, C., Brastianos, P. K., Brat, D. J., ... GLASS Consortium. (2019). Longitudinal molecular trajectories of diffuse glioma in adults. Nature, 576(7785), 112–120. https://doi.org/10.1038/s41586-019-1775-1

4. Kraboth, Z., & Kalman, B. (2019). Longitudinal Characteristics of Glioblastoma in Genome-Wide Studies. Pathology Oncology Research: POR. https://doi.org/10.1007/s12253-019-00705-1

5. Scarpace, L., Mikkelsen, T., Cha, Soonmee, Rao, S., Tekchandani, S., Gutman, D., ... Pierce, L. J. (2016). Radiology Data from The Cancer Genome Atlas Glioblastoma Multiforme [TCGA-GBM] collection. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9

6. Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery, 143(3), 331–336. https://doi.org/10.1016/j.otohns.2010.05.007

7. Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: Multivariate data analysis–an introduction to concepts and methods. British Journal of Cancer, 89(3), 431–436. https://doi.org/10.1038/sj.bjc.6601119