

Using Patient Attributes to Predict Glioblastoma Multiforme Survival Rates

Katie Harbaugh, Nathan Lam

July 31, 2020

Abstract

Glioblastoma is essentially one of the most lethal cancers of all, with a bleak prognosis and no determined cure. Additionally, the survival estimates given to patients are often quite inaccurate, as they can range from a few months to a few years based on multiple variables. The goal of this project is to test and create a model that relies on patient attributes to output a more specific and accurate survival curve. Through the use of Python packages, we were able to find correlations between certain characteristics and overall survival time. Then, by utilizing the Cox Model, we were able to output survival curves based on selected attributes, and compare the prediction with the results in the data. Given the time constraint, there is definitely more potential for fine-tuning this model than what was able to be achieved in 6 weeks. That being said, this project was successful in allowing us to create and test a prediction model that outputs a more specific percentage of survival at different intervals.

1 Introduction

Glioblastoma Multiforme (GBM) is a stage IV brain cancer, which is incredibly aggressive with few treatment options. The bleak prognosis for GBM patients is a several month life expectancy [5], partially due to the need for a combination of intense treatments to try and treat the tumor. Treatment often begins with surgical resection, but due to the complex invasive nature of the tumor, there is no way to fully remove the mass through surgery. So, it is often followed by a combination of therapies including radiation, chemo, etc. Even so, the purpose of these existing treatments is often with the realistic goal of prolonging life by a few months, as that seems to be as optimistic as possible.

My goal with this project is to analyze existing data sets that incorporate both specific attributes and survival rates for patients. In doing so, I hope to recognize correlations between the variables, and create an algorithm that improves prediction accuracy for patient survival. Although technology today is as advanced as ever, there are some inconsistencies in statistical output, especially when it comes to brain cancer. As mentioned, the average life expectancy is just over a year (about 14-15 months), but the reality is that patient survival is all over the board. Some patients who appear to have optimistic prospects end up dying after two months, while others live for 3 years[4]. Hopefully, by creating a program that more accurately predicts life expectancy, we can not only provide doctors with a timeline to help treat a patient more efficiently, but also provide families with a prediction that is more accurate, so they don't have to live with their loved one wondering if each day will be their last. However, it is important to note that exact survival predictions (i.e. 15.25 months) are extremely difficult and unrealistic. So, this project focuses more so on producing a survival curve for individual patients. Overall, our goals for this project are surely optimistic, but I feel that it is important to have objectives that are not just for the sake of furthering science, but maybe more importantly, for the sake of helping people who are affected by GBM each day.

2 Literature Review

Regarding the literature that I have come across so far, there seems to be a fair amount of information related to usage of the Cox Model for timeline predictions. While most of these publications are in the form of background information, documentation, etc., gaining as much knowledge as possible about the model has been very beneficial to my understanding of its framework. In the paper, Survival Analysis: Intuition & Implementation in Python, Anurag Pandey explains the background and usage of the Cox Model, while providing examples and code snippets for reference [8]. Without this paper, it would have taken much longer for us to understand the implementation of the model. Another paper that was very helpful in demonstrating attribute analysis for GBM is titled, Comparing Predictive Models of Glioblastoma Multiforme Built Using Multi-Institutional and Local Data Sources[7]. This paper's scope was more related to assessing the significance of certain variables on GBM survival, which helped give us an idea of which attributes to implement in our model. In this project we will take the concept a step further by producing survival curves based on this knowledge. The data sets I have selected to work with have an abundance of general information from hundreds of patients, so I am confident about the ability to conduct such research. Also, it is important to note that while Figure 1 below is not as simple as genetics/patient attributes to life expectancy, it is still immensely interesting to examine (See description below). To better understand the role of certain genes in relation to cancer, I did some more research relating to tumor suppressor genes, because in my data sets I noticed that the most commonly mutated genes were marked as "tumor suppressor genes." Upon doing further research, I learned that tumor suppressor genes are – when fully-functioning – meant to help regulate routine cell death and production [3]. However, when they mutate, they are obviously unable to carry out these processes/suppress tumors, allowing for un-monitored cell division and tumor growth. While learning this information was helpful to my

general understanding of Glioblastoma, it is less related to the overall scope of this project, and more related to my process of understanding the ins and outs of these tumors. Due to time and technology limitations, it was more realistic to analyze the given attributes, some of which being genetics-specific, and others not.

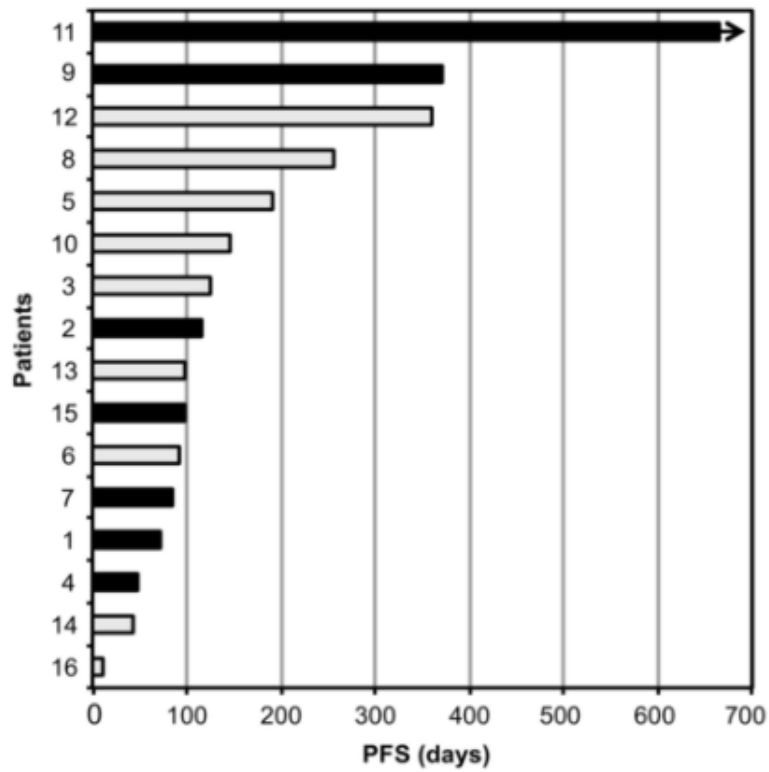


Figure 1: This graph depicts the PFS (Progression-Free Survival) in days for patients after GBM tumor resection. By analyzing the predominantly mutated tumor genes, and presenting the data before a tumor board, the board was able to recommend a treatment combination for patients. The bars in black represent those patients who opted in favor of the recommendation and proceeded with said treatment. The bars in grey represent patients which opted against the treatment recommended by the tumor board. Source: [5]

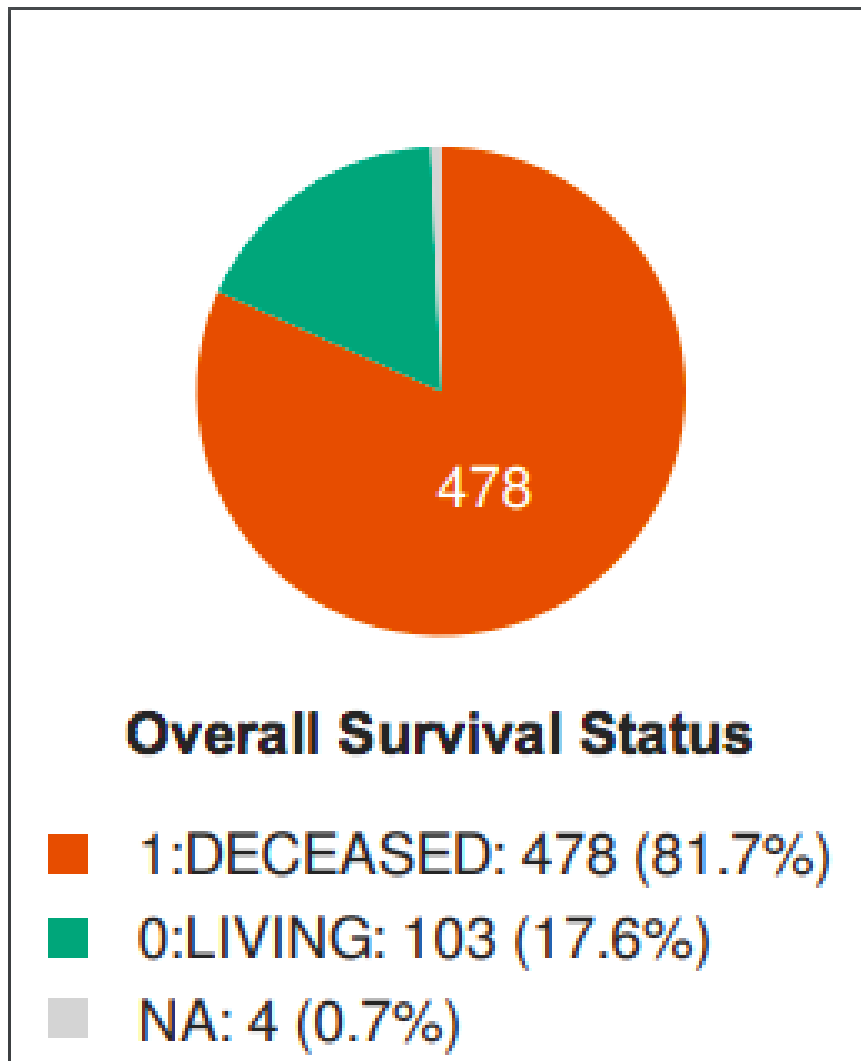


Figure 2: This figure stuck out for several reasons. First of all, although it may be basic, I thought it was important to include an image that represented the nature of GBM. With over 500 individuals in this trial, and 81.7 percent of them being deceased, it is obviously important to come up with a more accurate way of predicting this outcome ahead of time. One main reason for this is because if this data was estimated before its occurrence, it is very likely that treatment would have occurred at a different pace, frequency, etc. Source: [4]

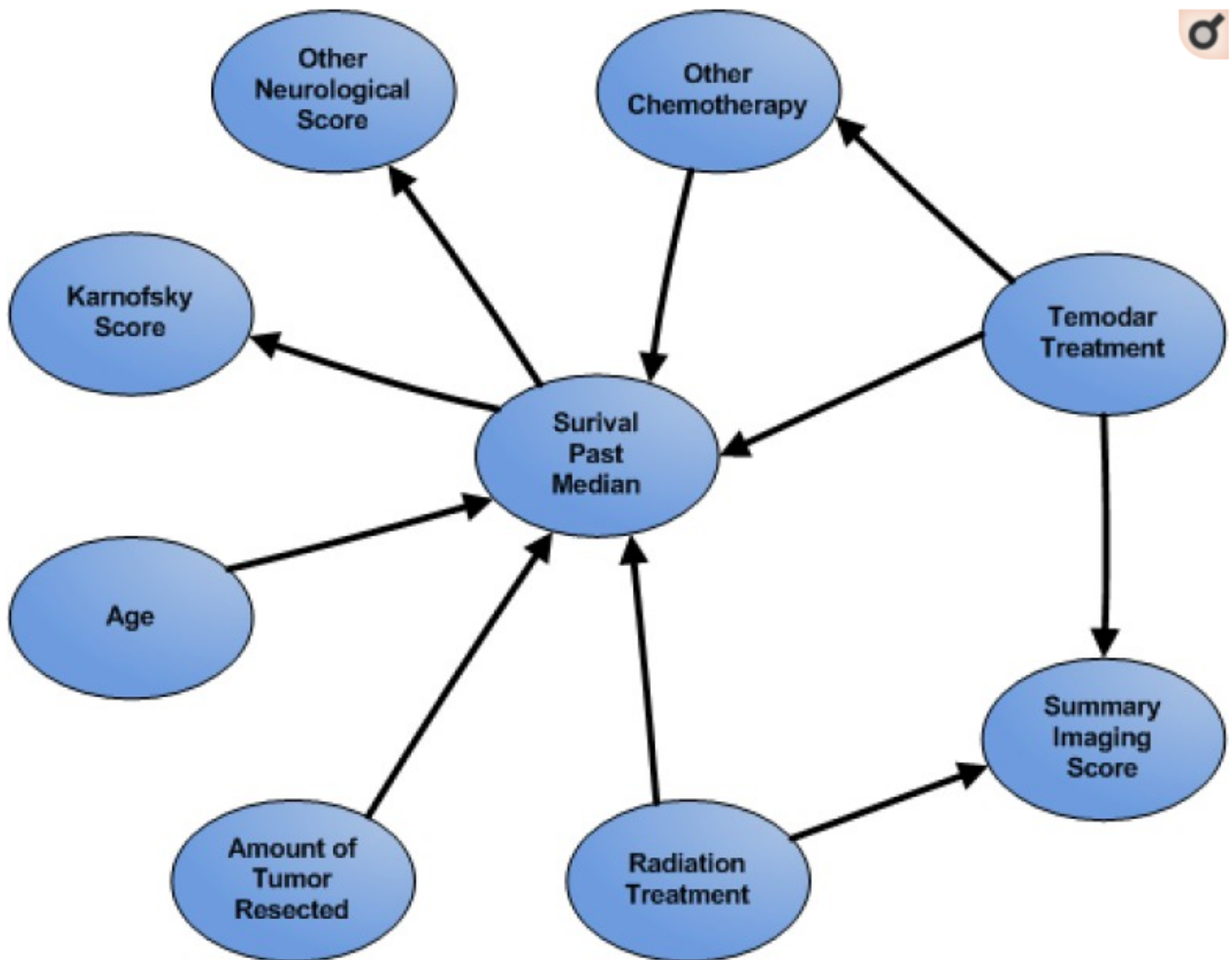


Figure 3: This diagram shows the connection between certain attributes and survival in GBM patients. For example, variables such as age, tumor resection, radiation treatment, etc. play a part in patient “Survival Past Median.” There are also many other connections in the diagram, such as how that survival affects tests such as the Karnofsky score, which makes complete sense. Not only is this a good visual representation of attribute relationships, but it will also be helpful when creating the algorithm and choosing which variables to take in. Source: [7]

3 Purpose

1. Goal 1: After further research, I have a good idea of which attributes are most important in predicting life expectancy, and the research-related goal of this paper is mainly to implement specific data and attributes for the sake of creating a more accurate prediction for survival at certain time intervals. More specifically, being able to do this through the use of technology I have access to at home. While this goal proved attainable within the time constraints, it is important to remember the big goal of this project, which was mentioned in the introduction.
2. Goal 2: Without a determined cure, the life of a Glioblastoma patient is quite bleak, knowing that there is pretty much only one way for things to end. To make this worse, survival estimates are given as overall percentages, rather than accurate, patient-specific values. So, the major goal of this project is to not only provide doctors with a more accurate timeline to follow for patient treatment, but more importantly to provide a more secure quality of life for individuals with this tumor and their families. To put it simply, if a more accurate prediction can help patients not have to live every day like it is their last, then this project is more than successful.
3. Goal 3: Finally, an underlying goal of this project was to become familiar with the process of conducting research and writing a research paper, regardless of the topic. After six weeks of doing so, I can comfortably say that I am far more confident about this process than I was at the beginning of the program, which I can mark as a success.

4 Methods

First of all, we have used Google Colab to conduct the programming required to complete this project. Additionally, I came across this statistical model called the Cox Model, which takes in specific patient attributes and helps output survival statistics.

The Cox Model outputs a "survival" curve, which can be used in many situations where time is affected by a number of variables. With the help of my mentor, I was able to practice using the Cox Model for the first time. One part of this was the implementation of dummy variables, which is very important in this situation because the model only reads quantitative values. Attributes such as age, survival in months, etc. are applicable here, but there are other variables we want to incorporate that are not numerical. Gender (is male) and overall survival status (deceased) were the selected variables in this instance, and were given values of "1" and "0" based on if the attribute for the patient was true or false, respectively [8]. I was also able to run a correlation on 5 selected variables (for simplicity) and the results were very interesting and provided decent insight into which variables are most negatively or positively correlated, especially with overall survival time. This is represented by Figure number 4, which can be seen below with a more detailed description.

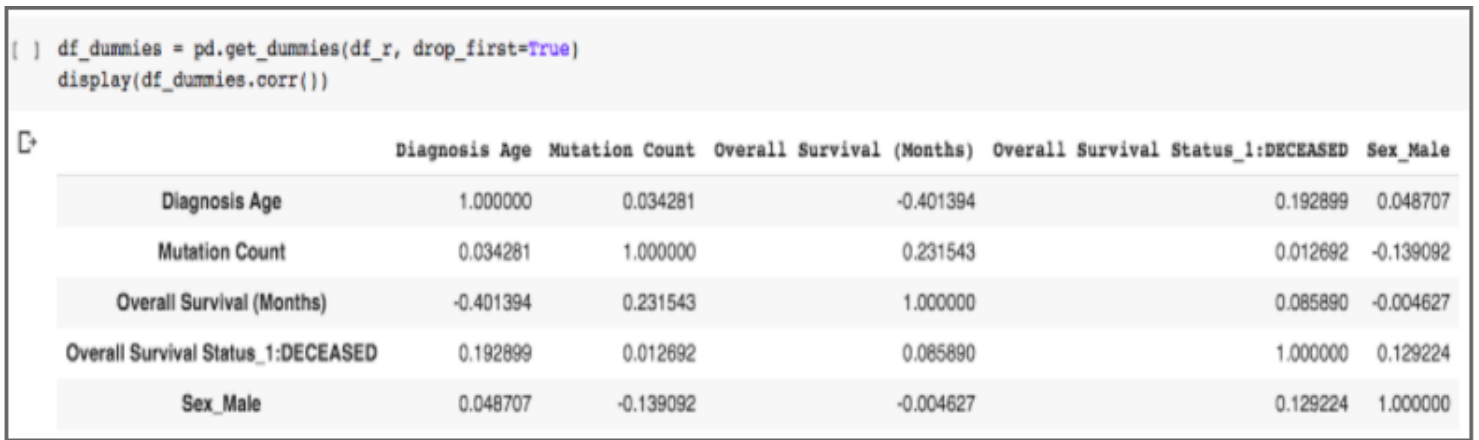


Figure 4: This is a code snippet from one of the programs we practiced implementing this week, based on data from: [1]. After running *get_dummies* on the selected variables to give quantitative values to qualitative variables (sex, overall survival status), we then displayed the correlation shown above, in a 5 by 5 table. The program works because, obviously, each column has one box with a value of “1.0” to represent a variable’s correlation to itself. Knowledge proven/gained by this test consists of how “Diagnosis Age” and “Overall Survival (Months)” are negatively correlated. This makes sense because a person is more likely to live longer if they are younger at diagnosis.

The main method used in this project was the Cox Model in the Lifelines package [6]. After displaying the data, running get dummies, and examining correlation, we then produced multiple survival curves based on different variables and data sets. Below, two examples of these regressions can be seen with more detailed descriptions.

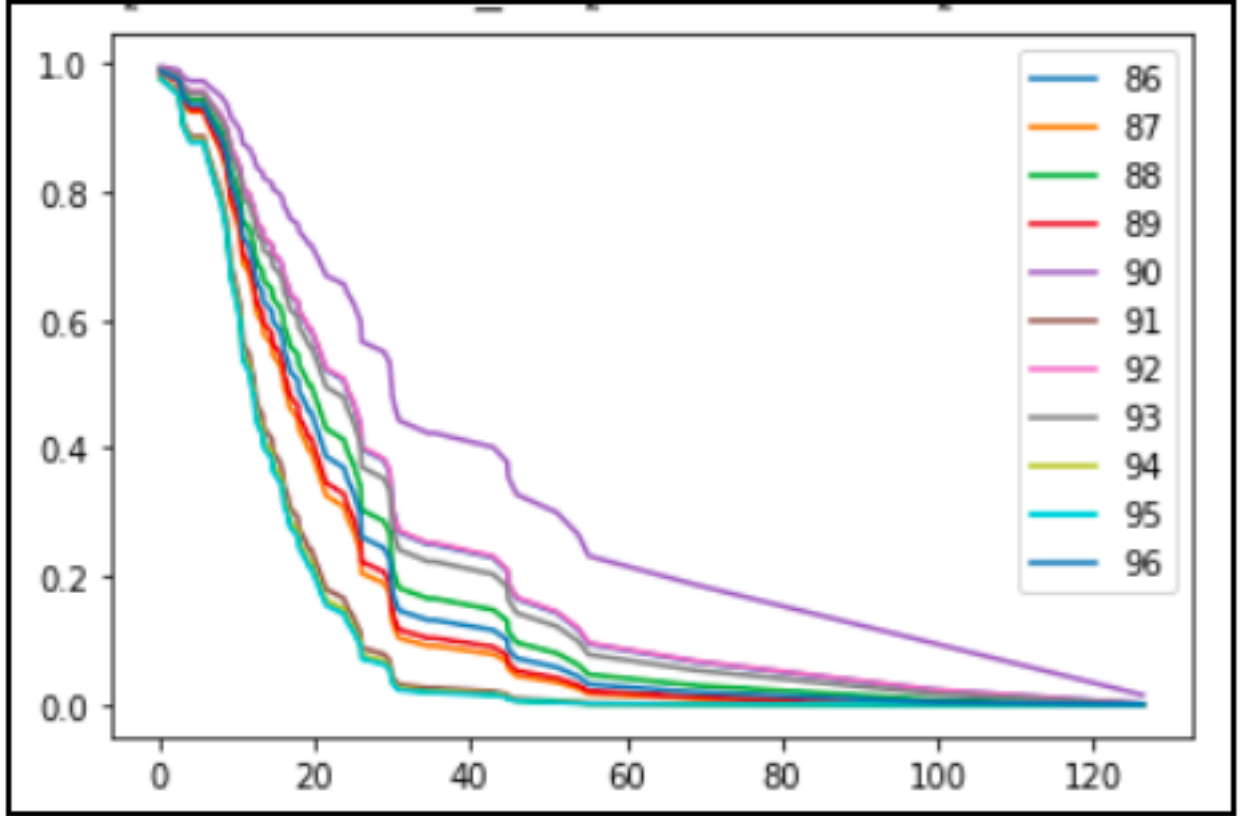


Figure 5: This figure shows survival curve predictions for the dataset Brain Tumor PDXs, 2019 [1]. The variables used to predict these curves were diagnosis age, sex, and mutation count. After cleaning the data of NaN (“not a number”) values, we split the samples into two sets: train and test. The test group consisted of 70, and the test set included the remaining 11. Each of the colored lines in the key represents an individual patient’s predicted survival curve, with the Y-axis being percentage of survival, and the X-axis being predicted survival in months

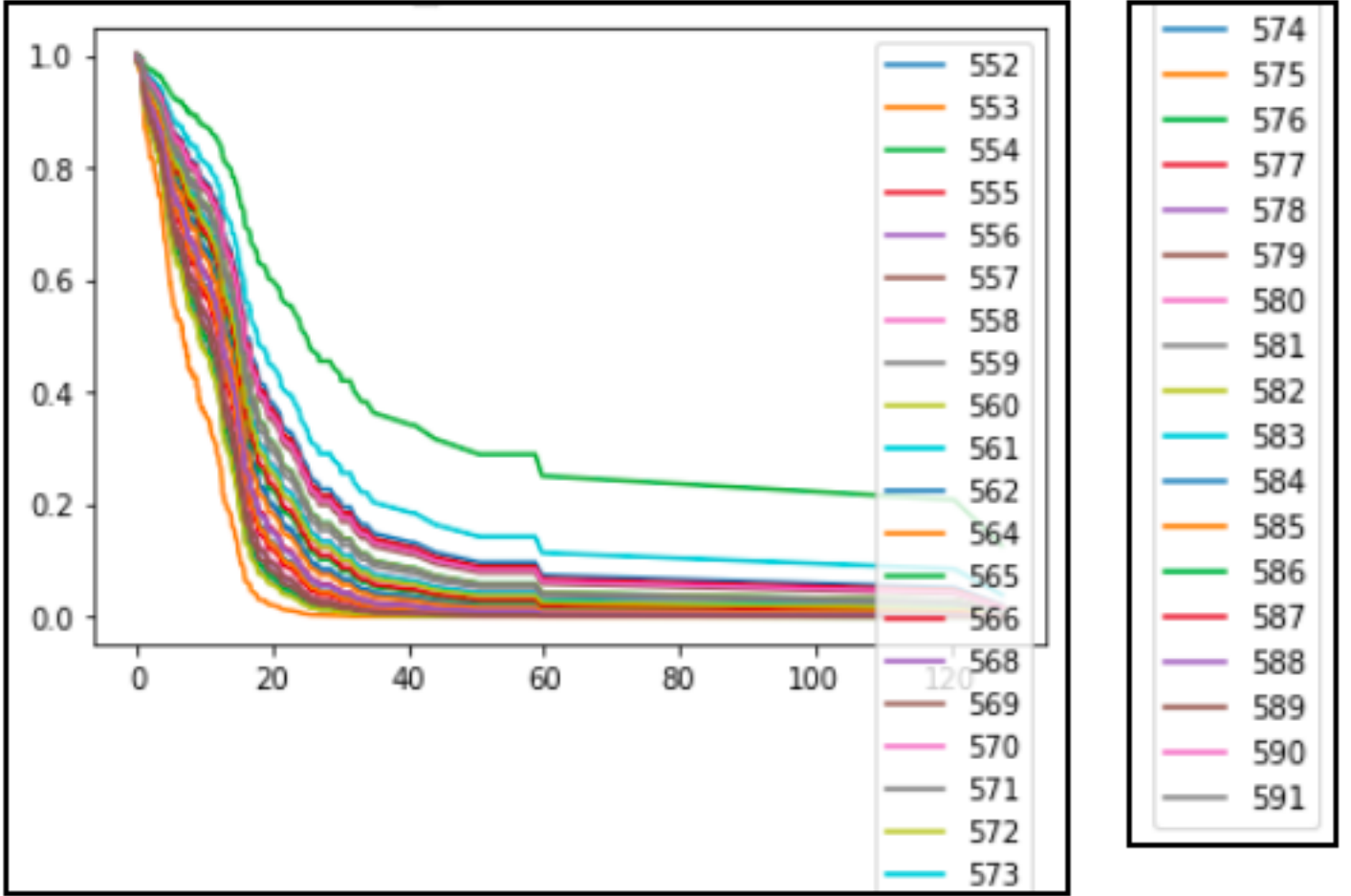


Figure 6: This figure is very similar to figure 5, but there are a few differences. First of all, the dataset used is the Glioblastoma Multiforme (TCGA, PanCancer Atlas), (2018)[4]. Also, while the event column and duration column (Overall Survival Status and Overall Survival (in months), respectively) are the same, the variables are not identical. Here, we used Diagnosis Age, Aneuploidy Score, and Fraction Genome Altered as our patient attributes. Finally, the dataset being much larger gave us a train set of 250 and a test set of 38, which explains the extensive key.

5 Results and Discussion

After working with our model and testing new variables and data sets, in the hopes of improving concordance accuracy, we were eventually able to increase the score tremendously by making some small changes. To put it simply, concordance is a metric used in this model to measure "goodness of fit" [9]. We discovered that the variable, "Disease Free (Months)," is highly correlated to overall survival time. This makes sense, given that if an individual's tumor is not progressing for said amount of time, it would be expected that their survival would increase accordingly. So, given that data set Glioblastoma Multiforme (TCGA, Firehose Legacy) [2] had this attribute for patient data, we decided to add it to the selected variables. By using a new data set and five selected attributes, the concordance score increased to 0.76, which represents a large improvement from the score of previously produces survival curves shown above, which had scores of roughly 0.66. An example of this curve for an individual patient can be seen in figure 7, with a more elaborate description. One important drawback to address is the the variable, "Disease Free (Months)" was mainly responsible for the concordance increase. So, if this data is not available in certain cases, the survival curve's prediction is expected to have a lower accuracy. That being said, with more time and testing, variables that have similar correlation strengths could be found, making the results less dependent on an individual attribute.

Additionally, we were able to visually analyze the summary output, regarding the final survival curve. Most of the values shown below in figure 8 serve the purpose of indicating the model's statistical significance. (All of the following information can be referenced in figure 8). For example, if a value in column "p" is particularly low, such as "Disease Free (Months)," it is very significant when it comes to concordance. Similarly, "Karnofsky Performance Score," "Sex_Male," and even "Diagnosis Age" all have low p values. However, we can also see that "Mutation Count" has a p value of 0.96, which means it is not significant to

the model's outcome, and could be removed without drastically altering the results. This is interesting, because one could think that the number of mutations a patient possesses would be negatively correlated to survival time. Based on general intuition, it would be reasonable to expect Glioblastoma patients with a large number of tumor mutations to have a low survival outcome. Nevertheless, our statistics reveal that this variable has a very low significance.

Similar conclusions can also be drawn from the column labeled "coeff" (for coefficient), where "Mutation Count" has a coefficient of 0.00, while "Sex_Male" has a coefficient of 0.53. The column to the right, "exp(coeff)" further explains these values. With an unaffected value of 1.00 (as seen for "Mutation Count"), this column shows the increase/decrease in likelihood of the event column's occurrence, when the specific covariate increases by 1. For example, when a patient is disease free for an additional month, their likelihood of death is decreased by 11%. This explanation is represented by "Disease Free (Months)" having an "exp(coeff)" of 0.89 (which is -11 from 1.00).

To the left of column "p," a column labeled "z" (for z score) can be seen. A score is given to each of the selected variables, which is a good indication of how significant each one is. Since z score is relative, we can see the impact of certain variables on the prediction. For example, "Disease Free (Months)" is again proven to be incredibly significant in determining life expectancy, as it has an extreme z score of -5.25. In similarly extreme ways, "Karnofsky Performance Score," has a z score of -2.60, while "Sex_Male," has a positive z score of 2.08. By means of simple visual analysis, our observations regarding each variable appear to be confirmed in each column of the model summary.

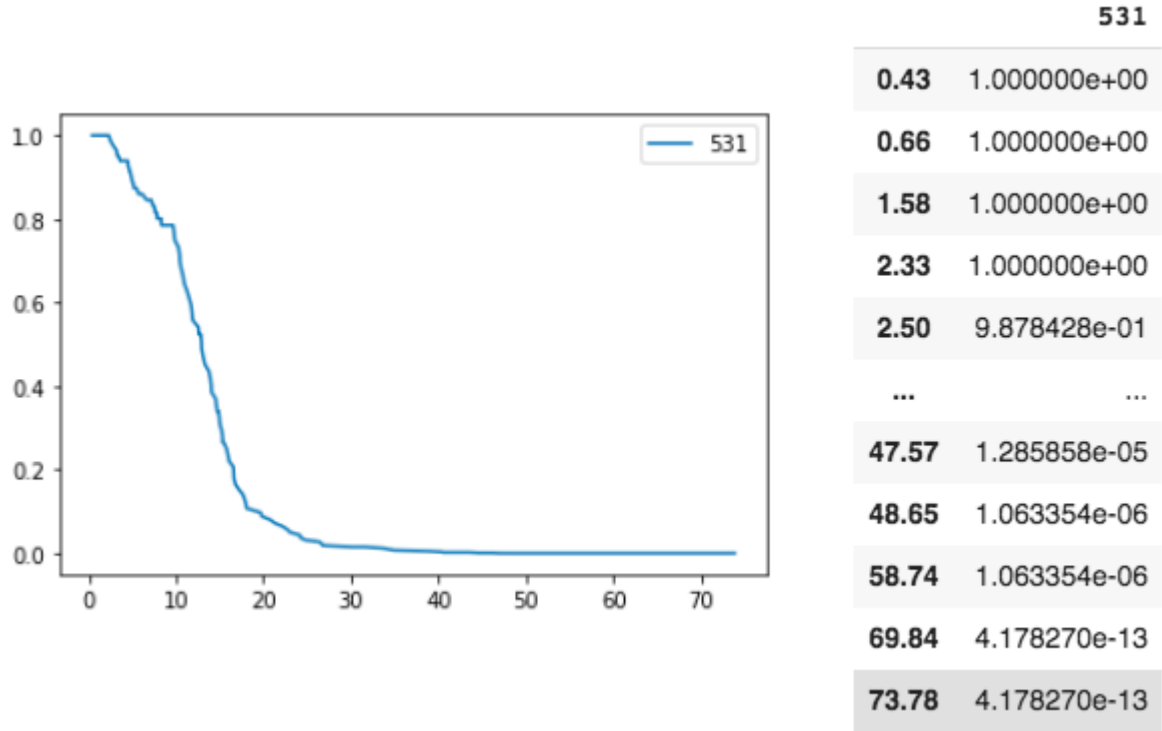


Figure 7: This figure shows the predicted survival for an individual patient in the dataset, Glioblastoma Multiforme (TCGA, Firehose Legacy) [2]. We used the same methods described in the methods section, but made some changes to the variables used. Here, we used these variables: "Diagnosis Age", "Disease Free (Months)", "Karnofsky Performance Score", "Mutation Count", and "Sex". The event and duration columns (Overall Survival Status and Overall Survival (in months), respectively) remained the same. On the right, you can see the numerical descriptions depicted in the patients survival curve. For example, the first row gives the patient a 100 percent chance of surviving 0.43-2.33 months. As time progresses, the percentage of survival decreases. At 2.5 months, the patient has roughly a 98 percent chance of survival, and an incredibly low chance of survival past about 40 months.

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
Diagnosis Age	0.02	1.02	0.01	-0.00	0.04	1.00	1.04	1.75	0.08	3.63
Disease Free (Months)	-0.12	0.89	0.02	-0.16	-0.07	0.85	0.93	-5.25	<0.005	22.68
Karnofsky Performance Score	-0.02	0.98	0.01	-0.04	-0.01	0.96	0.99	-2.60	0.01	6.76
Mutation Count	0.00	1.00	0.01	-0.02	0.02	0.98	1.02	0.05	0.96	0.06
Sex_Male	0.53	1.69	0.25	0.03	1.02	1.03	2.78	2.08	0.04	4.73

Figure 8: This is a screenshot of the printed summary for the model. Deeper descriptions for certain columns can be seen in the written Results and Discussion section, mainly for the columns which proved to be especially telling when understanding the results.

6 Conclusion

Although the time limit for this project was quite short, I feel that we were able to achieve solid results. We ended with a survival prediction model with a 0.76 concordance score, which I see as a huge success. We learned that certain variables are incredibly significant when it comes to accurately predicting survival, such as "Disease Free (Months)." Not only that, but we also found that certain attributes have virtually no significance at all (in this case), such as "Mutation Count." I am certain that with more time, this model's accuracy could be improved even further. That being said, we were able to create a model that produced individual survival curves, and a table for the individual curve detailing the percentage of survival over time. One future goal of this study is to condense our program code into its own package, or at least into an open-source format. In doing so, other individuals would get to implement their own patient data and variables to predict patient survival curves on their own. Obviously, there is plenty of work to be done between now and then, in terms of fine-tuning and model improvement, but this is still a realistic future direction to keep in mind. Another direction would be to refer back to some of my initial, more genetics-oriented ideas of using patient-specific genetic data to produce survival curves. It would be interesting to see how this information could po-

tentially affect the concordance score, or the overall accuracy. Again, this would require much more time than allotted for this project, but it is definitely worth consideration as a future project.

References

- [1] Brain tumor pdxs. <https://www.cbioportal.org/study/clinicalData?id=gbm_mayo_{pdxs}arkaria₂₀₁₉>.
- [2] Glioblastoma multiforme (tcga, firehose legacy). <https://www.cbioportal.org/study/summary?id=gbm_tcga>.
- [3] Oncogenes and tumor suppressor genes. Available at <https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/oncogenes-tumor-suppressor-genes.html>.
- [4] Glioblastoma multiforme (tcga, pancancer atlas). Available at <https://www.cbioportal.org/study/clinicalData?id=gbm_tcga_{pancanceratlas}2018>, 2018.
- [5] Sara A. Byron, Nhan L. Tran, Rebecca F. Halperin, Joanna J. Phillips, John G. Kuhn, John F. de Groot, Howard Colman, Keith L. Ligon, Patrick Y. Wen, Timothy F. Cloughesy, Ingo K. Mellinghoff, Nicholas A. Butowski, Jennie W. Taylor, Jennifer L. Clarke, Susan M. Chang, Mitchel S. Berger, Annette M. Molinaro, Gerald M. Maggiora, Sen Peng, Sara Nasser, Winnie S. Liang, Jeffrey M. Trent, Michael E. Berens, John D. Carpten, David W. Craig, and Michael D. Prados. Prospective feasibility trial for genomics-informed treatment in recurrent and progressive glioblastoma. *American Association for Cancer Research*, 2017.
- [6] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, sean reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel Wilson, Gabriel, Luis Moneda, Kyle Stark, Arturo Moncada-Torres, Harsh Gadgil, Jona, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, Dave Golland, jlim13, and Abraham Flaxman. Camdavidsonpilon/lifelines: v0.25.0, July 2020.
- [7] BSE Kyle W. Singleton, PhD William Hsu, and PhD Alex AT Bui. Comparing predictive models of glioblastoma multiforme built using multi-

institutional and local data sources. *National Center for Biotechnology Information*, 2012.

- [8] Anurag Pandey. Survival analysis: Intuition + implementation in python. *Towards Data Science*, 2020.
- [9] Elizabeth Atkinson Terry Therneau. Concordance. *National Center for Biotechnology Information*, 2020.