"Lung Cancer Trial Survival Analysis"

IEE 573: Reliability Engineering Term Project

Nishan Shetty

Abstract

Reliability analyses can be applied to medical clinical trials to determine the effectiveness of a novel treatment. The Veteran's Administration Lung Cancer Trial dataset was reviewed to 1) determine variable relationships; 2) compare the reliability of the standard and test chemotherapy treatment; 3) predict survivability of the patients with lifetime distributions, the Cox proportional hazards model, random survival forest, and Cox multilayer perceptron; and 4) determine individualized treatment recommendations for patients. Weibull distribution was found to best explain the lifetime distribution the best using goodness of fit measures like AIC, -2Loglikelihood, BIC and Anderson Darling. The novel treatment showed to have higher reliability than the standardized treatment but was not always recommended due to a patient's characteristics. It was found that the Karnofsky score and histological cell type were significant predictive drivers throughout all models. The Cox proportional hazards model had the highest concordance index value (0.763); however, this model is likely invalid for the dataset. The Cox multilayer perceptron resulted in a c-index of 0.73, while random survival forest had the lowest c-index value.

Table of Contents

Table of Figures

1.0 Introduction

Clinical trials assess the effectiveness of novel medical approaches (e.g. strategy, treatment, device) to address a specific health condition amongst individuals. There are innate risks involved with participating in a clinical trial, but the entire process is managed and monitored by a team of professionals to protect the participants. Clinical trials may be a reliability measurement of the proposed medical approach. In the Veteran's Administration Lung Cancer Trial, a survival analysis was conducted to compare the efficacy between a standard and test chemotherapy for advanced, inoperable lung cancer across the four cell types: small, adenocarcinoma, squamous, and large (**Table 1**). Each of these cancerous lung cancer cells needs to be treated differently to achieve effective results.

Cell Type	Percent of Cases	Description
Small	15	Most aggressive and rapidly growing; typically caused by tobacco; can spread throughout the body quickly
Adenocarcinoma	40	Most common form; located on the peripheral areas of the lungs; prevalent in non-smoking women
Squamous	30	Located in the central lung area in the bronchi, mostly localized; generally linked to smoking
Large	15	Least common type; grows throughout the lung; commonly spreads to lymph nodes and distant sites

Table 1. Types of Lung Cancer

The purpose of this study is to analyze the reliability of the test chemotherapy treatment by the following methodology:

1. Perform initial exploratory data analysis to understand underlying relationships amongst the dataset,
2. Conduct survival time analysis to determine the reliability of the clinical trial treatment compared to the standard treatment with lifetime distributions,
3. Compute and compare predictive models (Cox proportional hazards, random survival forest, and Cox Multilayer Perceptron (MLP) by their concordance index (c-index), which can be used to find concordant pairs, i.e., if the risk of the event predicted by a model is lower for the patient who experiences the event later point, and
4. Generate personalized recommended treatment procedures for each patient.

By applying standard and predictive survival analysis methodologies to the clinical trial data, the team seeks to determine which if the test treatment was effective and if one predictive model has a significantly more accurate response amongst the tested models.

2.0 Data Exploration

The Veteran's Administration Lung Cancer Trial was a clinical trial utilized to evaluate the effectiveness of new chemotherapy for advanced, inoperable lung cancer (Kalbfleisch, 1980). There were 137 patients who participated in the study with six variables of interest recorded, with only 9 of the survival time being censored:

1. *Survival Time* - exact survival time of the patients.
2. *Status* - categorical variable (yes/no) indicating whether the survival time was censored.
3. *Treatment* - categorical variable depicting rather the patient received the test chemotherapy or standard chemotherapy throughout the clinical trial.
4. *Histological Type of Tumor* - categorical variable of the tumor type (i.e. squamous cell, small cell, adenocarcinoma cell, or large cell).
5. *Diagnosis* - a continuous measure of the time (in months) from when the patient was diagnosed to the start of the clinical trial.
6. *Age* - patient's age in years.
7. *Prior Therapy* - categorical variable (yes/no) stating whether the patient had received any prior therapy to the start of the clinical trial.
8. *Karnofsky Performance Scale Index* - a ranked measure of patients' performance status, where a lower value indicates a lower likelihood of survival (**Table 2**).

Able to carry on normal activity and work with no special care necessary	100	Normal activity; no evidence of disease
	90	Normal activity; minor signs/symptoms of the disease
	80	Normal activity with effort; some signs/symptoms of the disease
Unable to work but can live at home and care for most personal needs with varying levels of assistance	70	Independent but cannot carry out a normal activity or active work
	60	Requires occasional assistance but can care for day-to-day needs
	50	Requires significant assistance and medical care
Unable to care for self; requires institution or hospital care; disease may be progressing rapidly	40	Disabled; requires significant assistance and medical care
	30	Severely disabled; hospital admission indicated; death not imminent

	20	Very sick, hospital admission necessary; active supportive treatment mandatory
	10	Dying; fatal processes progressing rapidly
	0	Dead

Table 2. Karnofsky Performance Status Scale Definitions Rating (%) Criteria (Abernethy, 2005)

There are a few limitations to clinical trials. First, clinical trials are performed on human subjects. There are innate differences of characteristics to experiences from individual to individual, which makes it impossible to restrict the success or failure of a novel treatment to the treatment alone. Second, individuals who are willing to participate in a clinical trial may not be representative of the population. Third, participation in a trial may impact the results as individuals may take better care of themselves during the time frame. Lastly, exogenous factors such as quality of life or satisfaction, which are not frequently measured during clinical trials, may have significant impacts (Collet, 2000). While not all these limitations can be adjusted for, data exploration can help recognize any confounding or correlated variables prior to data analyzation.

First, the normality of the dataset (**Figure 1**) was examined, and it was determined that the data was right-skewed. This relation remained true while examining the subsections of categorical variables for prior therapy and cell type. These histograms may be viewed in Appendix A. After checking the normality assumption was examined, individual variables were analyzed for correlations. Generally, a correlation factor of greater than 0.8 is considered of concern, and this value was not achieved across the full dataset or any subsections of data. The strongest relationship appeared between the Karnofsky score and survival time with the strongest relation shown for the adenocarcinoma cell subgroup (**Figure 2**). It is surprising that the correlation between Karnofsky score and survival time is not stronger given that the Karnofsky score seeks to determine the health of the patient; however, this scoring index is qualitative and subjective, which could impact the categorizations patients receive.

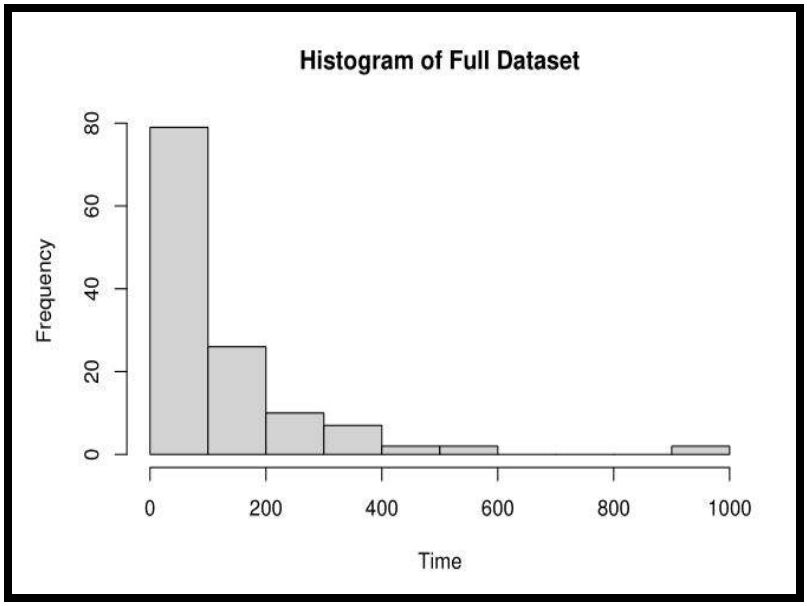


Figure 1: Histogram depicting survival time for the entire dataset

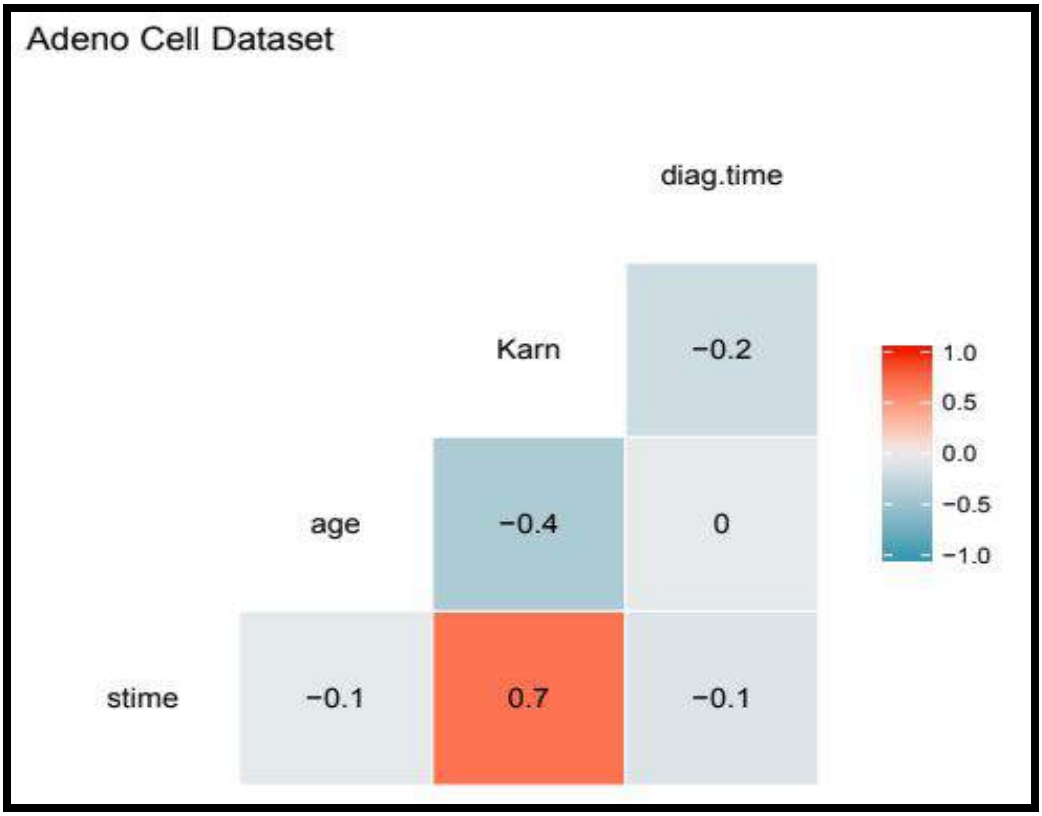


Figure 2: Correlation plot for continuous and ranked (%) variables

Lastly, the Kaplan-Meier estimation was used to compute survival functions for various subgroups of the dataset to explore the relationships between the variables and patient survival.

These computations were completed within R (an open source statistical software) utilizing the survival package. The R Markdown file that contains the data cleaning, normality assumption check, correlation plots, and survival functions may be reviewed in Appendix B. The full dataset survival curve may be viewed in **Figure 3**. The following comparative observations were made amongst the subgroups of variables:

- *Histological Type of Tumor* - Patients with small cell lung cancer had the longest survival times followed by large cell, squamous cell, and adenocarcinoma cell (**Figure 4**).
- *Treatment* - Patients receiving the test chemotherapy had a longer survival time than those who did not (**Figure 5**).
- *Prior Therapy* - Patients who had received prior therapy had a longer survival time than those who did not (**Figure 6**).
- *Age* - With the definition of elderly being set as 65 years old, patients who were elderly had shorter survival times than those who were not classified as elderly (**Figure 7**).
- *Karnofsky Performance Scale Index* - Patients with a Karnofsky score of less than or equal to 40 (unable to care for oneself) had shorter survival times than those with higher Karnofsky scores (**Figure 8**).

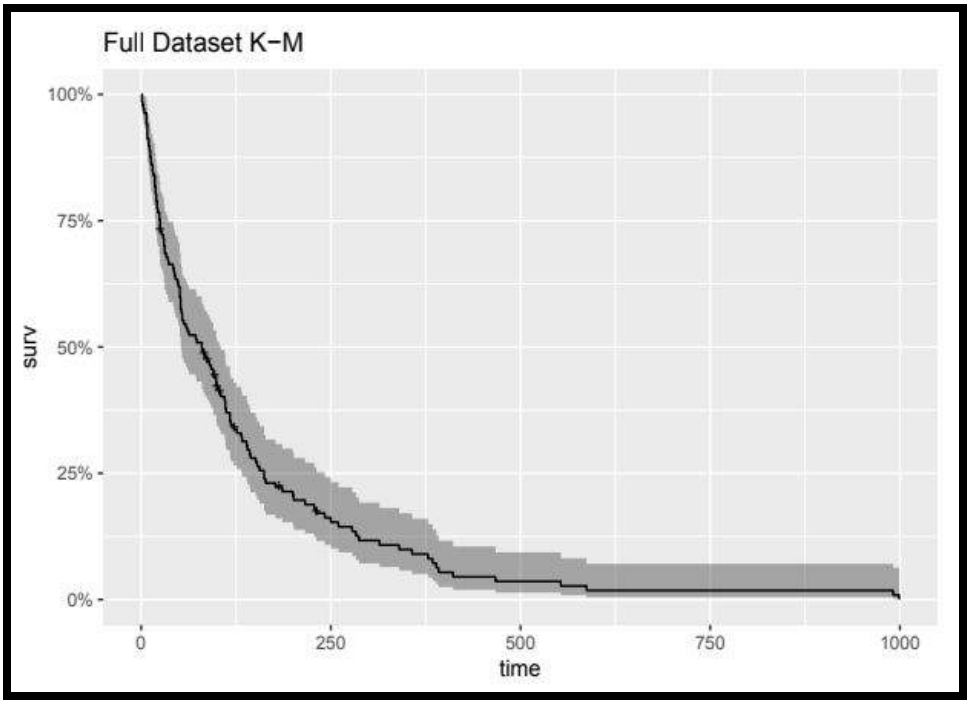


Figure 3: Survival plot representing the entire dataset

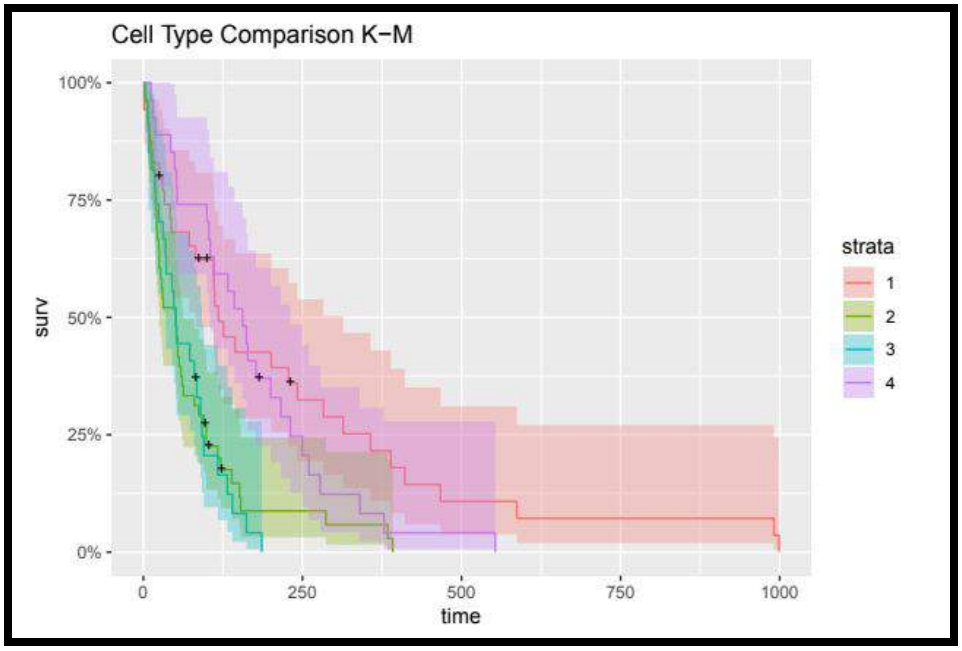


Figure 4: Survival plot of the entire dataset classified by histological type of tumor with the following labels: 1 = small cell, 2 = squamous cell, 3 = adenocarcinoma cell, and 4 = large cell

.

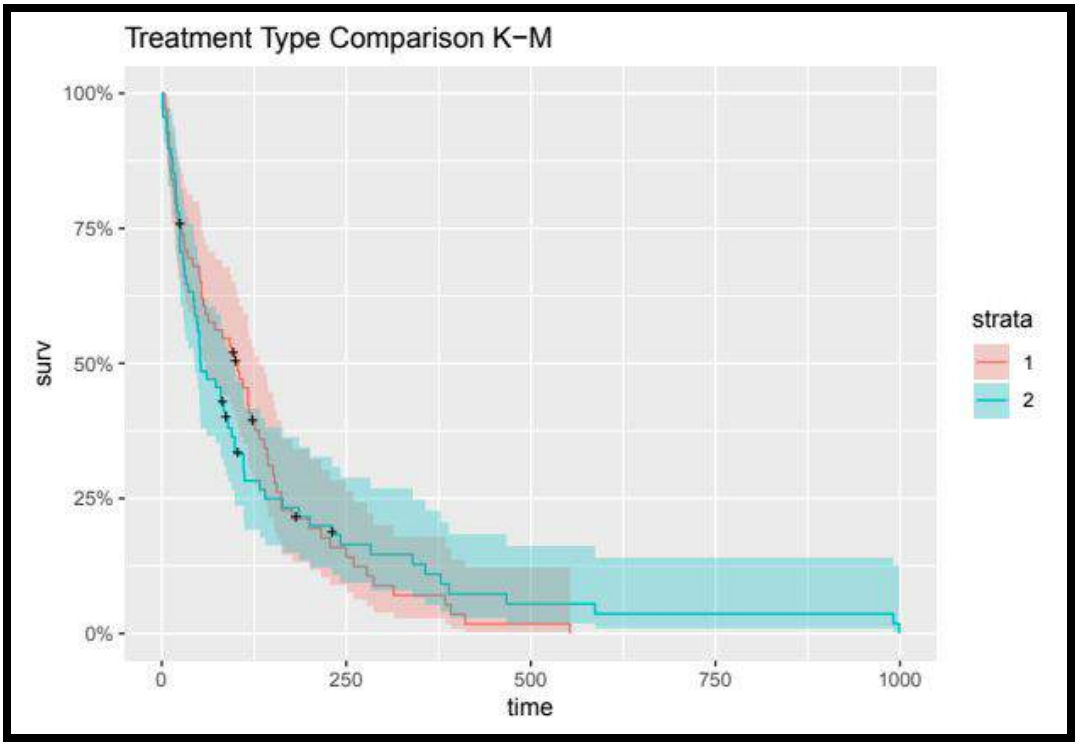


Figure 5: Survival plot of the entire dataset classified by treatment type with the following labels: 1 = standard and 2 = test

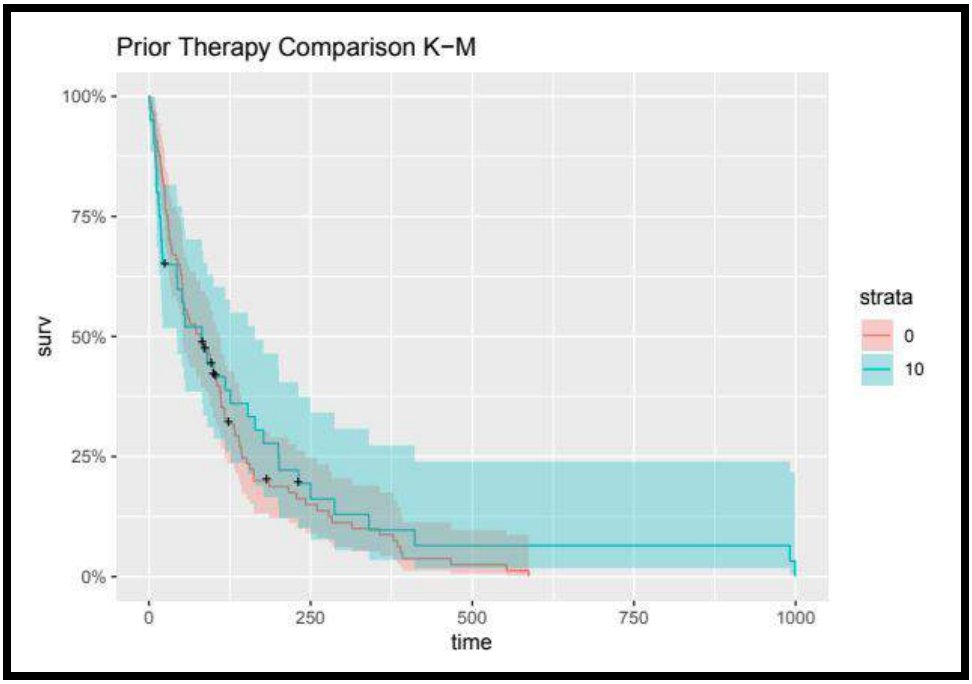


Figure 6: Survival plot of the entire dataset classified by prior therapy with the following labels: 0 = no and 2 = yes

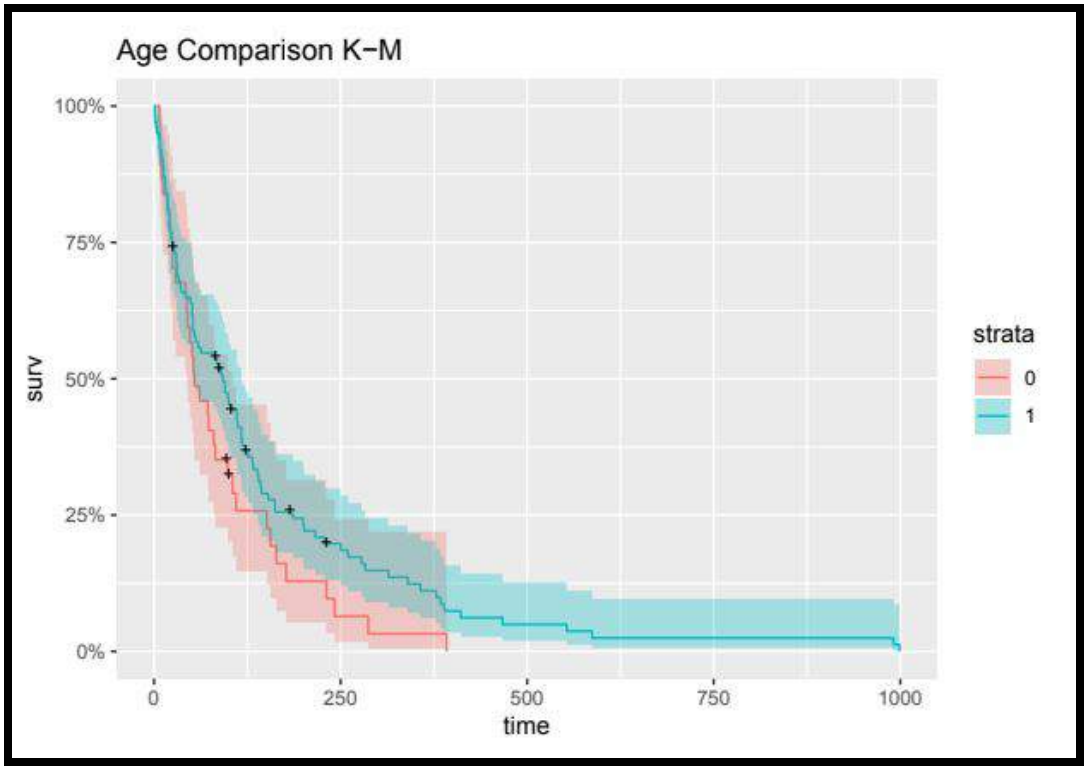


Figure 7: Survival plot of the entire dataset classified by age with the following labels: 0 = 65+ years and 1 = less than 65 years

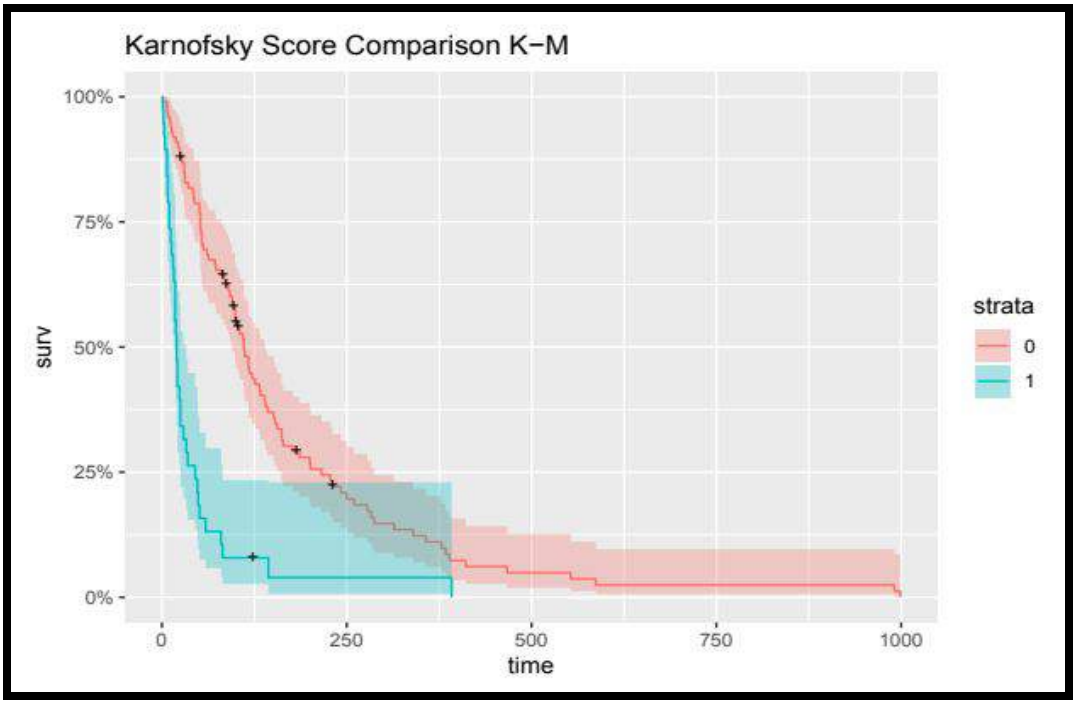


Figure 8: Survival plot of the entire dataset classified by Karnofsky score with the following labels: 0 = 50 to 60 and 1 = 0 to 40

.

This data analysis showed no strong correlations between dependent variables, but noticeable differences in the survival time, or the response variable, when manipulating sub-populations. Initial conclusions suggest that the Karnofsky score will be a dominant predictor in the survival of the patient. Since there was not a strong correlation between the Karnofsky score and either the standard or test treatment sub-populations, the longer survival time observed for the test treatment patient appears to be attributable to the novel treatment. The following sections will introduce survival analysis methods and the results of applying these processes to the Veteran's Administration Lung Cancer Trial.

3.0 Methods

3.1 Lifetime Distributions

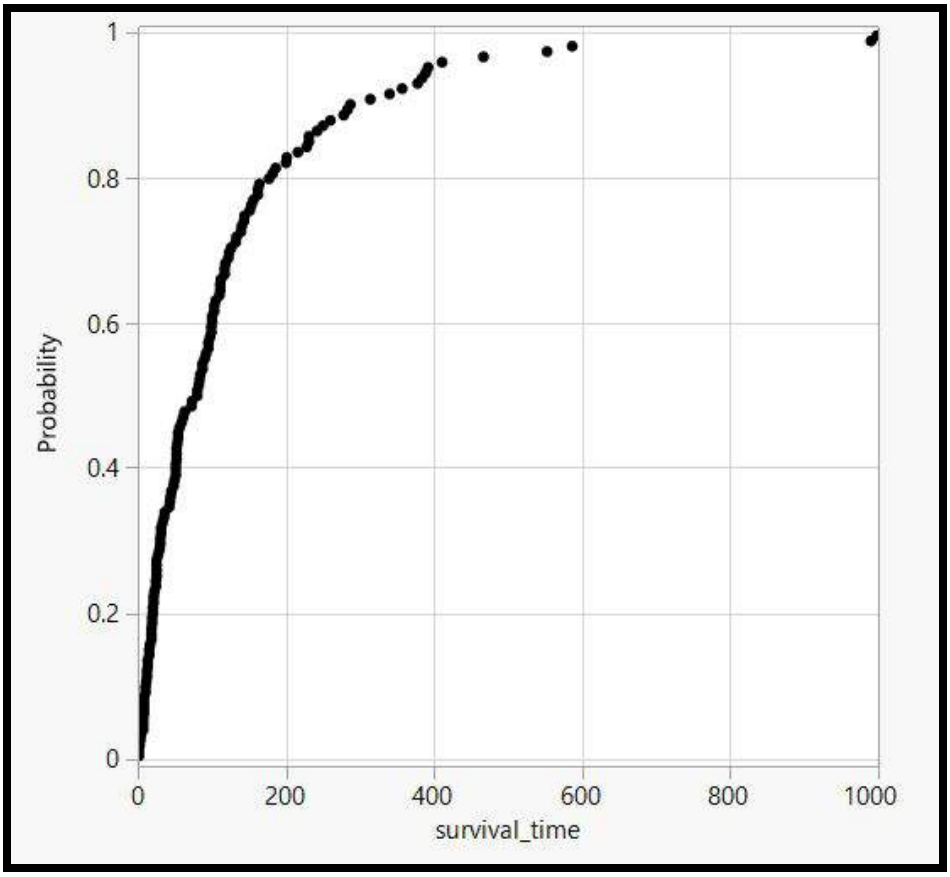


Figure 9: Lifetime distribution of survival time

The lifetime distribution is a collection survival time graphically presented as a plot as number of failures versus the survival time.

Distribution	AIC	-2Loglikelihood	BIC
Weibull	1588.2181	1584.1285	1593.9385
Lognormal	1591.1540	1587.0644	1596.9044
Exponential	1591.4941	1589.4644	1594.3844
Loglogistic	1592.1796	1588.0900	1597.9300
Fréchet	1635.1918	1631.1023	1640.9423
LEV	1669.7341	1665.6445	1675.4845

Logistic	1725.7287	1721.6392	1731.4791
Normal	1778.7081	1774.6185	1784.4585
SEV	1910.5188	1906.4293	1916.2692

Table 3. Comparisons of distributions for best fits

The highlighted distributions were selected to compare for the best fit of the survival time. Weibull, Lognormal and Exponential distributions show the best results according to AIC, -2Loglikelihood and BIC. Normal distribution was also selected for purposes of comparison, is a very commonly used parametric model.

This distribution is compared using the goodness of fit measures like Anderson Darling, AIC, BIC, and -2Loglikelihood to determine the best distributions among normal, Weibull, exponential and lognormal distribution.

Distribution	Probability Density Function	Parameters
Normal	$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$	μ = mean of the normal times-to-faiure, also noted as $\bar{T}$, θ = standard deviation of the times-to-failure
Weibull	$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$	η = scale parameter, or characteristic life β = shape parameter (or slope)
Exponential	$f(t) = \lambda e^{-\lambda t}$	λ = constant rate, in failures per unit of measurement,
Lognormal	$f(t') = \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t'-\mu'}{\sigma'}\right)^2}$	$t' = \ln(t)$. t values are the times-to-failure μ' = mean of the natural logarithms of the times-to-failure σ' = standard deviation of the natural logarithms of the times-to-failure

Table 4. Distributions with their pdfs

3.2 Cox Proportional Hazards Model

The Cox Proportional Hazards model (CPH) is a semiparametric technique utilized to fit survival curves for multiple predictors (MathWorks, 2019). The ability to calculate survival with an input of multiple predictors is the primary advantage of this method. It is a regression that does not assume a lifetime distribution but does assume the contributions of the predictors remain constant over time. This function may be used with both quantitative and qualitative variables. A hazard function is calculated as

$$h(X_i,t) = h_0(t)\exp[\Sigma x_{ij} b_j]$$

where X_i equals the various predictors (x) for each (i^{th}) patient, $h_0(t)$ represents the common baseline hazard function, t is time, and b_j is a coefficient that measures the effects of each predictor. Derived negative coefficients have a positive correlation with the survival curve while negative coefficients have a negative relation. The effect is utilized to calculate the estimated hazard ratio [$HR(X_i)$] of each predictor [$\exp(b_j)$] from the following equation:

$$HR(X_i) = [h(X_i,t)] / [h_0(t)] = \exp[\Sigma x_{ij} b_j]$$

The predictor variables are then compared with the baseline HR, producing

$$HR(X_i) = [h(X_i,t)] / [h_0(t)] = \exp[\Sigma (x_{ij} - x_j^*)\, bj]$$

where x_j is the mean value of the given predictor. This HR is then integrated into the survival rates by

$$Sx_i(t) = S_0(t)^{HR(Xi)}$$

where $S_0(t)$ is the survival function.

3.3 Random Survival Forest

Random survival forest (RSF) is a nonparametric, ensemble tree machine learning method for analysis of right-censored survival data of prospective cohorts where the outcome is a time-dependent variable. Logrank splitting rule is used.

RSF uses a collection of decision trees for prediction and to rank variables by their importance for time to the event, a method which has also been successfully applied to identify risk factors of different diseases. Variable importance is obtained using VIMP package in R.

The ensemble is constructed using tree-based Nelson-Aalen estimators, and the conditional cumulative hazard function is estimated below:

$$\hat{H}_b(t|\mathbf{x}) = \int_0^t \frac{N_b^*(\mathrm{d}s, \mathbf{x})}{\bar{Y}_b^*(s, \mathbf{x})}$$

The ensemble survival function is

$$\hat{S}^{\mathrm{rsf}}(t|\mathbf{x}) = \exp\left(-\frac{1}{B}\sum_{b=1}^{B} \hat{H}_b(t|\mathbf{x})\right)$$

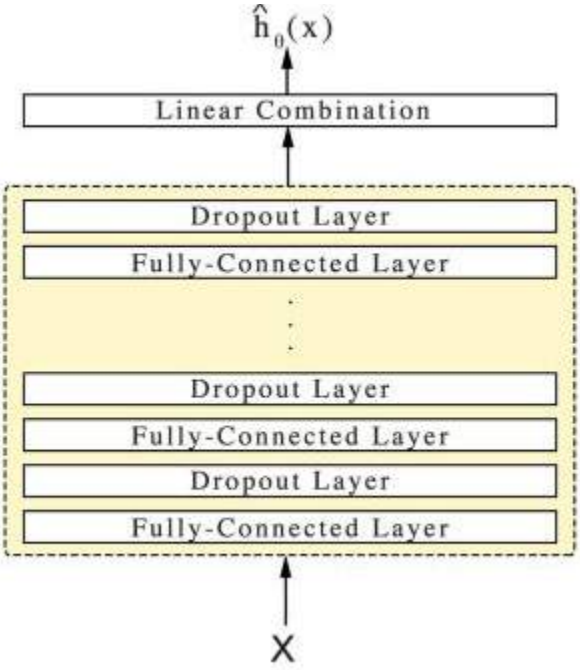
Where B is the number of bootstrap samples, $N_b^*(ds, x)$ counts the uncensored events until time ds, $Y_b^*(s, x)$ is the number of patients at risk at time s.

3.4 Cox Multilayer Perceptron

DeepSurv is a Cox proportional hazards deep neural network and state-of-the-art survival method for modeling interactions between a patient's covariates (eg: clinical and genetic

features) and effectiveness of various treatment options to provide personalized treatment recommendations. It can successfully model increasingly complex and nonlinear relationships between a patient's covariates and their risk of failure.

It is a multi-layer perceptron, which estimates the risk function $h_\theta(x)$ in a single node using the below network θ and input x:



The loss function is the L2-regularized negative log partial likelihood with weight decay shown in the equation below:

$$l(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j \in \Re(T_i)} e^{\hat{h}_\theta(x_j)} \right) + \lambda \cdot ||\theta||_2^2,$$

The regularization with weight decay along with the addition of dropout layers in the network will help in preventing the training process from overfitting.

In a clinical study, patients are subject to different levels of risk based on their relevant prognostic features and which treatment they undergo. In the context of our dataset, let $\tau = 1,2$ be the treatment groups where each treatment i has an independent risk function $h_i(x)$. For any patient, the network should be able to accurately predict the risk $h_i(x)$ of being prescribed a given treatment i. If we define the baseline hazard function $\lambda_0(t)$ as being equal for each patient, then we can take the log of the hazards ratio to calculate the personal risk of prescribing one treatment option over another. We define this difference of log hazards, or $rec_{ij}(x)$:

$$\mathrm{rec}_{ij}(x) = \log\left(\frac{\lambda(t;x|\tau=i)}{\lambda(t;x|\tau=j)}\right) = \log\left(\frac{\lambda_0(t)\cdot e^{h_i(x)}}{\lambda_0(t)\cdot e^{h_j(x)}}\right)$$
$$= h_i(x) - h_j(x).$$

4.0 Results

4.1 Lifetime Distributions

N	N*	Mean	StDev	Median	Minimum	Maximum	Skewness	Kurtosis
137	0	121.628	157.817	80	1	999	3.12659	13.0701

Table 5. Descriptive Statistics of the survival time

Table 5 shows the descriptive statistics of the survival time with 137 observations having a mean equal to 121.628 and standard deviation of 157.817. The range of survival times is from 1 to 999. The skewness and kurtosis were found to be 3.12659 and 13.0701 respectively.

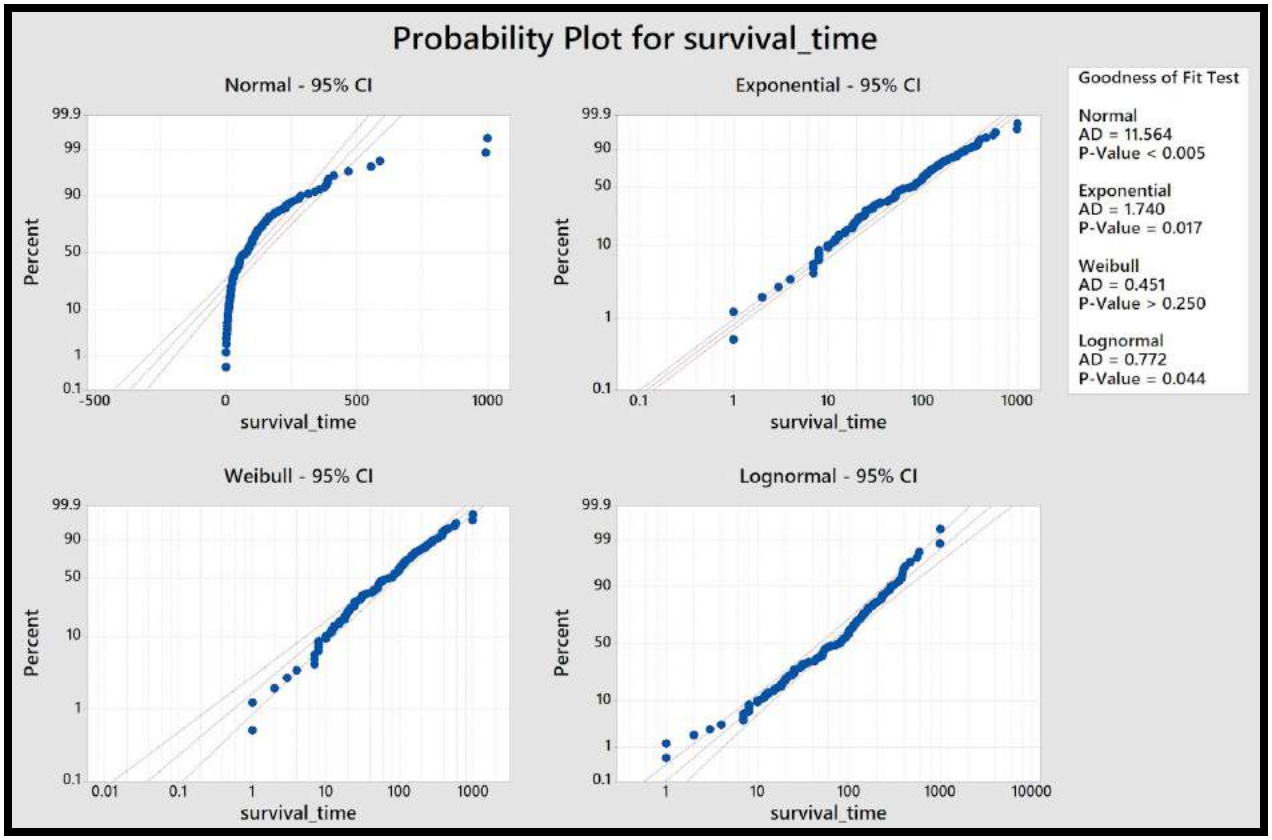


Figure 10: Comparison of distributions using Anderson Darling

In Figure 10, a comparison of the normal distribution, exponential distribution, Weibull distribution, and lognormal distribution is carried out using Anderson-Darling.

Distribution	AD	P
Normal	11.564	<0.005
Exponential	1.740	0.017
Weibull	0.451	>0.250
Lognormal	0.772	0.044

Table 6. Goodness of fit test of distributions using Anderson Darling

Distribution	**AIC**	**-2Loglikelihood**	**BIC**	**Anderson Darling**
Normal	1778.7081	1774.6185	1784.4585	11.564
Weibull	1588.2181	1584.1285	1594.8855	0.451
Exponential	1591.4941	1589.4644	1594.3844	1.740
Lognormal	1591.1540	1587.0644	1596.9044	0.772

Table 7. Comparisons of distribution using AIC, -2Loglikelihood, BIC and Anderson Darling

In Table 7, these distributions are compared using AIC. -2Loglikelihood, BIC and Anderson Darling, all these measures indicate that Weibull distribution and lognormal distribution are the two best distributions which show the best fits among the above four distributions.

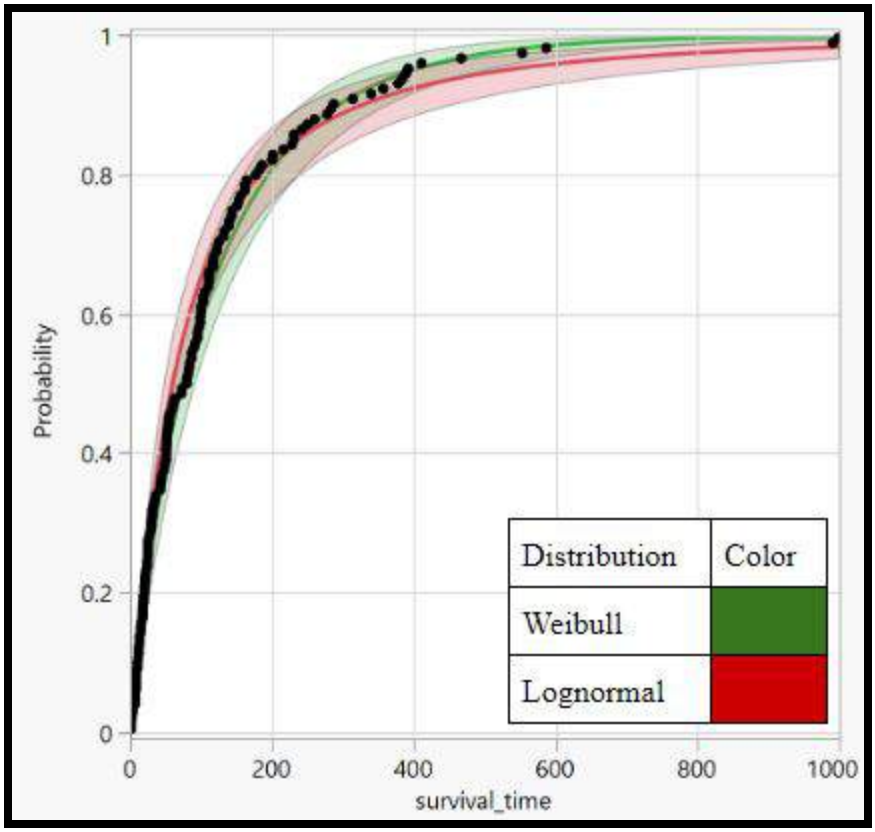


Figure 11: Weibull and lognormal distribution comparison

A graphical comparison (Figure 11) of Weibull and lognormal distribution is carried out to verify the best distribution which fits the survival time data. Looking at the probability plots and the parameters such as AIC, -2Loglikelihood, BIC, and Anderson Darling, it can be concluded that the Weibull distribution fits the survival data the best.

Parameter	Estimate	Std Error	Lower 95%	Upper 95%	Criterion	
location	4.72245	0.104151	4.518318	4.92658	-2*LogLikelihood	1584.1285
scale	1.15339	0.074186	1.007985	1.29879	AICc	1588.2181
Weibull α	112.44347	11.711100	91.681296	137.90745	BIC	1593.9685
Weibull β	0.86701	0.055766	0.769948	0.99208		
Mean	120.86265	11.923112	99.614141	146.64364		

Table 8. Parameter Estimates of Weibull Distribution

The parameter estimates of Weibull distribution were found out. Being a two-parameter Weibull distribution, the estimated value of location and scale were 4.72245 and 1.15339 respectively. The standard error along with lower 95% and upper 95% was also computed.

4.2 Cox Proportional Hazards Model

The significant variables were the Karnofsky score, squamous cell, and adenocarcinoma cell, and the coefficients and exponents may be viewed in **Table 9**. The Karnofsky score resulted in a negative coefficient, indicating that this predictor has a protective effect on the model. Meanwhile, the presence of squamous or adenocarcinoma cells had a positive coefficient, leading these predictors to be negatively correlated with survival. The data exploration supports these findings and shows no disruptions. The c-index of this model was computed to be 0.763 with a standard error of 0.030. The final fitted survival curve may be viewed in **Figure 12**, and the .Rmd file may be viewed in Appendix B.

Predictor	Coefficient	Exponent	Significance
Karnofsky Score	-0.0382	0.9677	0.001
Squamous Cell	0.8616	2.3670	0.01
Adenocarcinoma Cell	1.1960	3.3070	0.001

Table 9. CPH Significant Variables

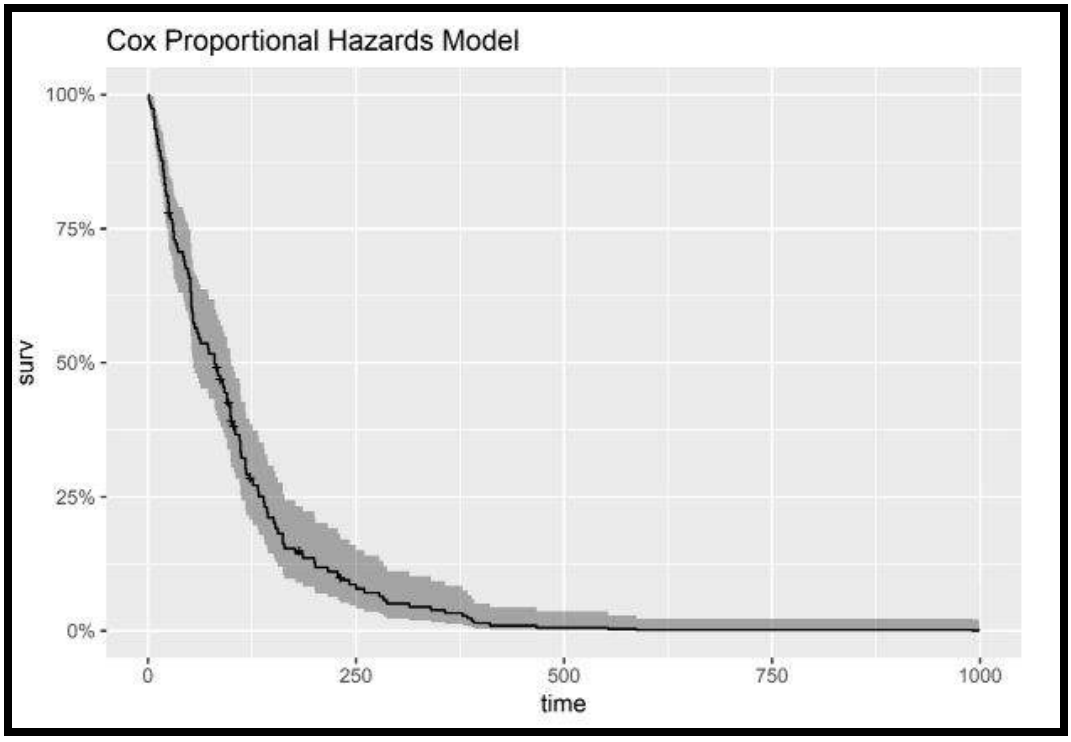


Figure 12: All predictors fitted to a survival curve

However, as aforementioned, this model operates under the assumption that the variables are not time dependent. When reviewing if this condition is satisfied, it is evident that the variables are not completely constant over time (**Figure 12**). Therefore, this assumption may not be

completely satisfied. Future research could explore holding the variables constant by subsetting the data over periods of time where the predictor does not change.

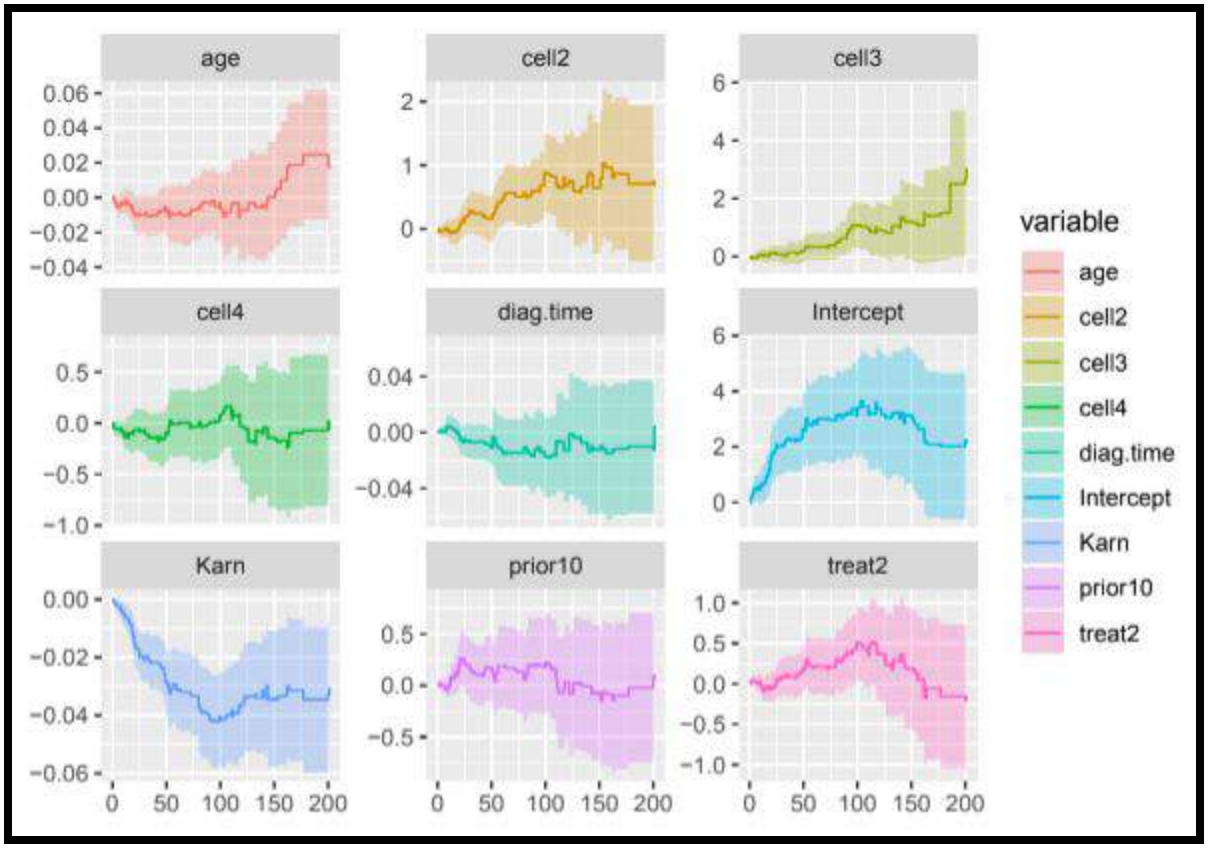


Figure 13: Time dependence of age, squamous cell, adenocarcinoma cell, large cell, diagnosis, Karnofsky score, prior therapy, and test treatment

4.3 Random Survival Forest

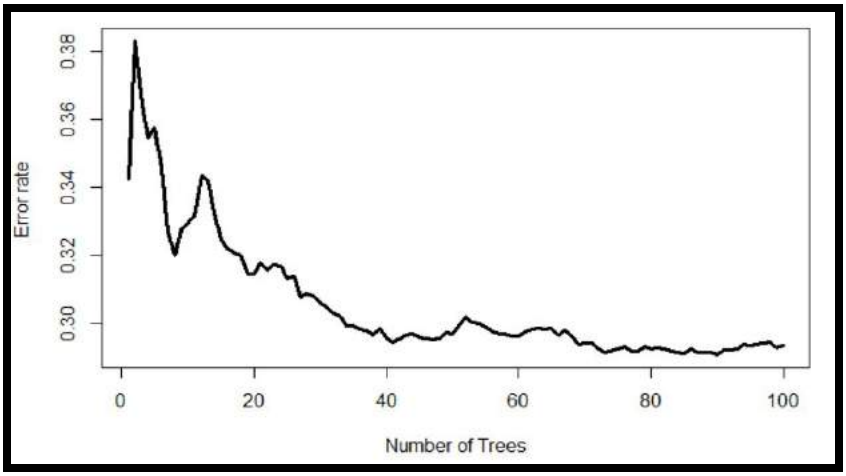


Figure 14: Learning curve over 100 Trees

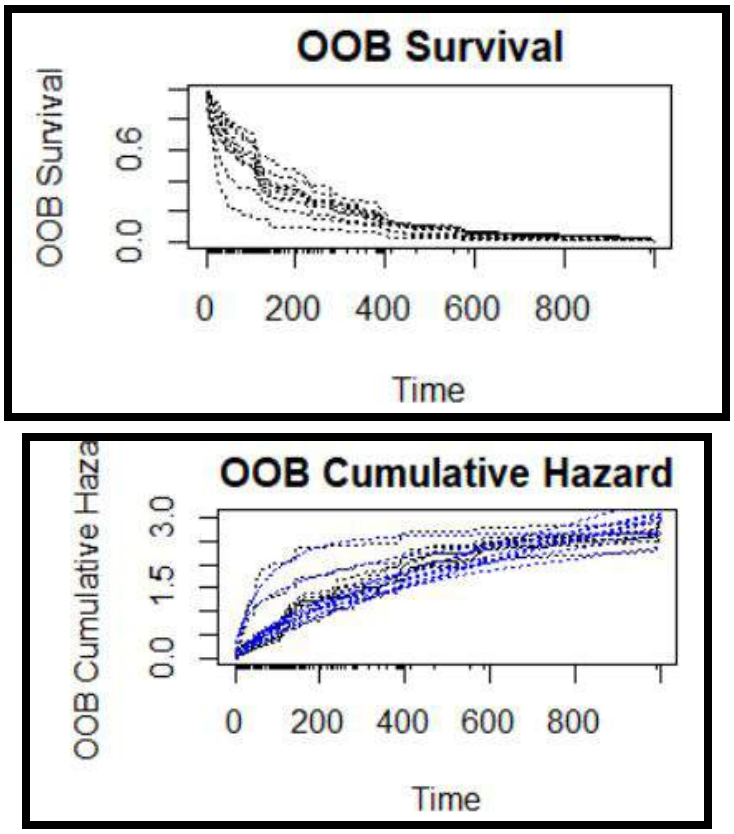


Figure 15: Survival and Cumulative Hazard Function for Out of Bag (OOB) data

The concordance index obtained was calculated to be 0.7, and the variable importance (VIMP) values were obtained for each variable along with the standardized values, as shown in **Table 10**.

Predictor	VIMP	Standardized VIMP

Treatment	241.20016	0.009684374
Cell Type	2900.13121	0.116442524
Status	65.63755	0.002635399
Karnofsky Score	5159.26693	0.207148580
Diagnosis Time	108.23530	0.004345732
Age	-103.51786	-0.004156323
Prior Therapy	187.89281	0.007544042

Table 10. Random Survival Forest Variables of Importance

From the values above, it is shown that the variables most important in the predictive ability of the model are "Karnofsky Score," "Cell Type," and "Treatment." The survival times predicted by those variables are plotted below:

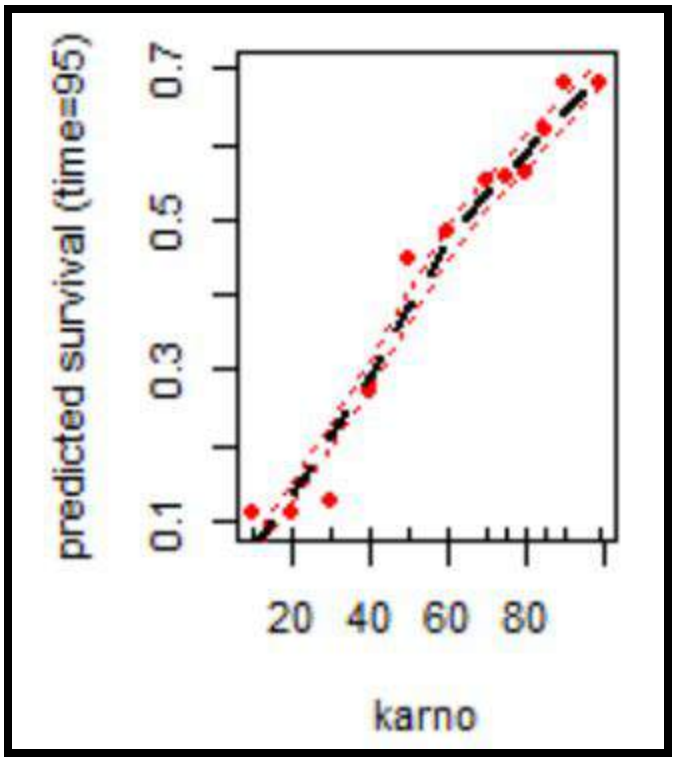


Figure 16: Predicted Survival by Karnofsky Score variable

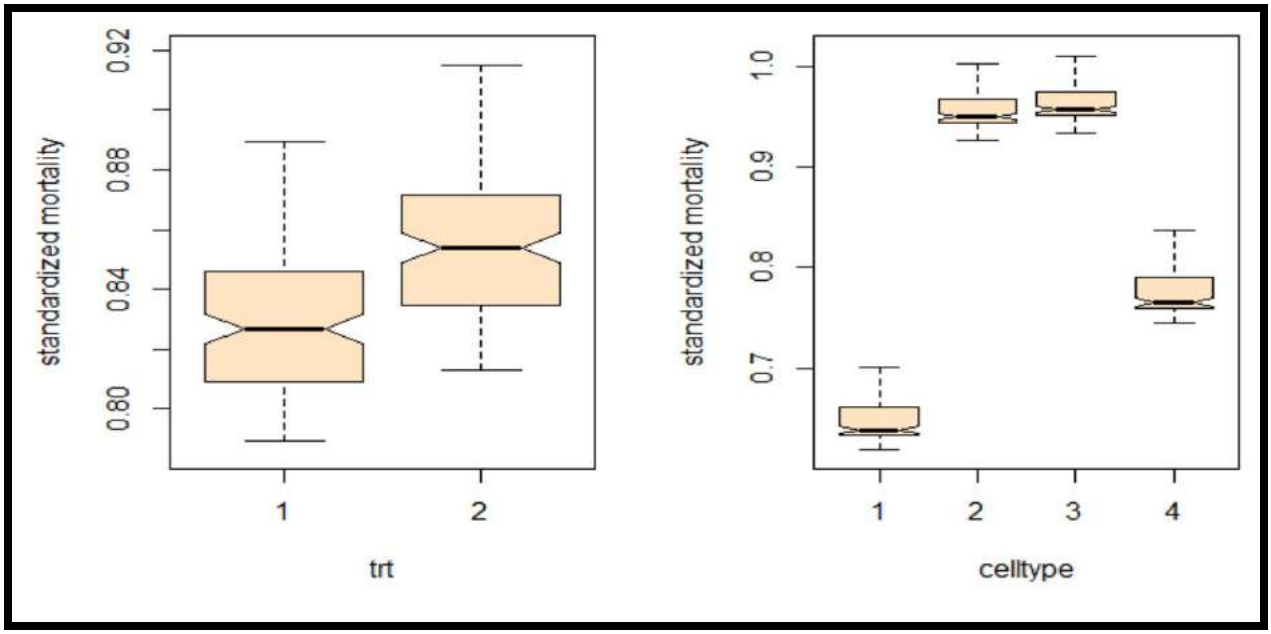


Figure 17: Predicted Survival by treatment variable (left), and Cell Type variable (right)

4.4 Cox Multilayer Perceptron

The model was trained for 2000 epochs, using the Adam optimizer and negative partial log-likelihood loss function to obtain the concordance indices or c-indices for the training and validation set for each epoch. The Training set c-index obtained is 0.73, and the validation set C-Index obtained is 0.71. Below are the learning curves for the loss function and c-index over each epoch:

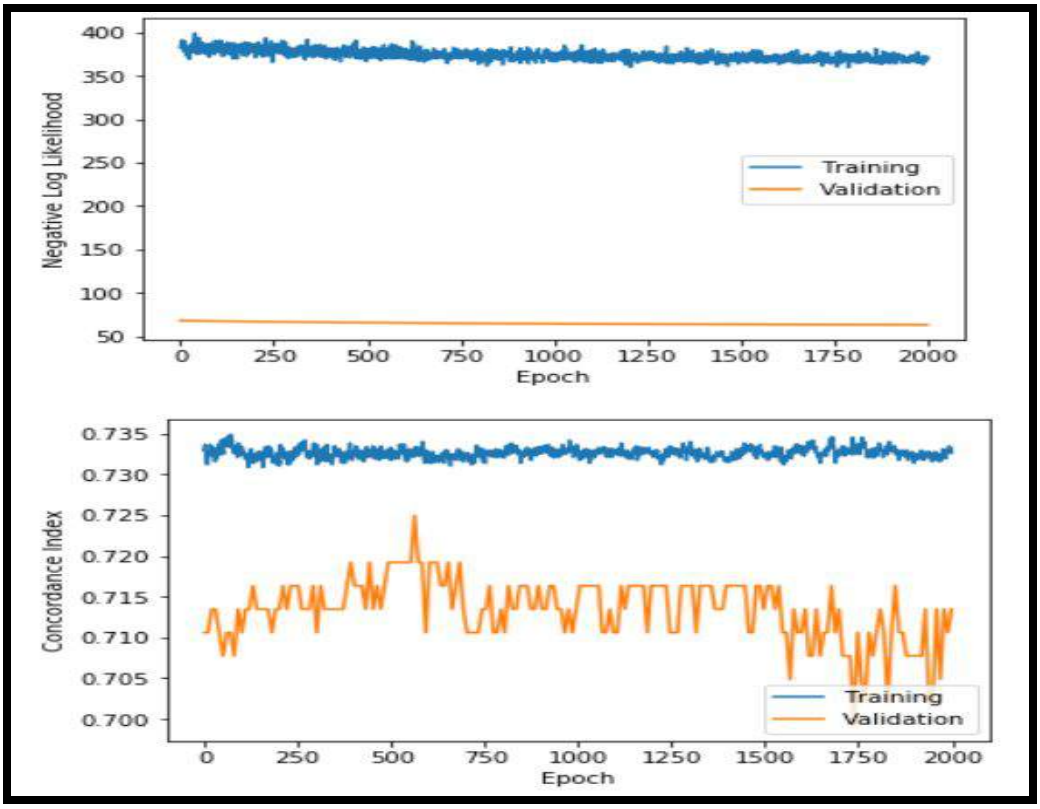


Figure 18: Learning curves for Cox multilayer perceptron for each epoch

Personalized treatment recommendations were calculated for each patient using the defined $rec_{ij}(x)$ function. A sample of these calculation results is visible in **Table 11**.

Patient	Coefficient	Treatment Recommendation
0	-0.869285	2
1	6.892474	1
2	2.344195	1
3	0.420172	1
4	0.501167	1
5	-0.189242	2
6	-0.870105	2
7	0.119706	1

8	4.271060	1
9	-0.150997	2

Table 11. Personalized Treatment Recommendations

As visible in **Table 11**, if the function computed for the function is negative, treatment 2 is recommended for the patient. Conversely, if the function computed for the function is positive, treatment 1 is recommended for the patient.

5.0 Discussion

First and foremost, the test treatment proved to increase survivability collectively of the patients as seen through fundamental lifetime distribution techniques. By taking this a step further, the study sought to create predictive models of patient survivability within the characteristic range. A comparison of the predictive models showed high predictability for Cox proportional hazards followed by Cox multilayer perceptron and random survival forest as seen in **Table 12**. However, the data inputted into the Cox proportional hazards model, as aforementioned, invalidates the time-dependence assumption. To validate the Cox proportional hazards model in future research, a stratified Cox proportional hazards model could be implemented to demonstrate the changes over time or the model could be simplified to only examine a period where the covariates remain constant.

Model	C-Index
Cox Proportional Hazards	0.763
Random Survival Forest	0.708
Cox Multilayer Perceptron	0.732

Table 12. Comparison of Predictive Models

Therefore, within the scope of this analysis, the Cox multilayer perceptron is the best predictive model. The advantage of the Cox MLP's architecture is that it does not require a priori specification of treatment interactions terms, unlike CPH which would compute a constant recommender function unless interaction terms are added to the model which can be time-consuming and expensive due to the required experimentation and biological knowledge needed. Further research should also be conducted exploring why the Cox models had a higher predictive capacity than the random survival forest model. It is believed that the performance of the random survival forest could be increased by further tuning the hyperparameters. K-fold cross-validation for the RSF and Cox MLP could potentially have them converge to the global optima of their loss functions, and thus potentially increasing their respective C-Indices. The assumption that all the predictor variables are time-invariant also likely does not hold true.

The most rewarding aspect of this study was the personalized treatment recommendations as this prediction could help automate the medical diagnosis performed by doctors to improve

individual patient survivability as the test treatment was not always recommended depending upon the patient's characteristics. This computational diagnosis combined with the test chemotherapy treatment can help increase the number of lives saved.

6.0 Conclusion

Lung cancer is the leading cause of cancer death in the US, with overall 5-year survival of approximately 16%. US veterans have a 76% higher incidence of lung cancer than the general population. The purposes of this project are to review, develop and compare different survival analysis models on right censored lung cancer clinical trial patients' data. Specifically, the Veteran's Administration Lung Cancer Trial dataset was reviewed to 1) determine variable relationships; 2) compare the reliability of the standard and test chemotherapy treatment; 3) predict survivability of the patients with lifetime distributions, Cox proportional hazards model, Random Survival Forest, and Cox Multilayer Perceptron; and 4) use the Cox Multilayer Perceptron determine personalized treatment recommendations for patients.

Weibull distribution was found to best explain the lifetime distribution the best using goodness of fit measures like AIC, -2Loglikelihood, BIC and Anderson Darling. It was found that the Karnofsky score and histological cell type were significant predictive drivers throughout all models. The Cox proportional hazards model had the highest concordance index value (0.763); however, this model is likely invalid for the dataset. The Cox Multilayer Perceptron resulted in a c-index of 0.73, while Random Survival Forest had the lowest c-index value.

The novel treatment has also been shown to generally lead to a higher survival rate compared to the standard treatment for the cohort but was not always individually recommended by the Cox MLP due to a patient's covariates (clinical and genetic features, among others). The function used to compute this is the difference between the log hazards ratio of both treatments for each patient.

7.0 References

Abernethy, A. P., Shelby-James, T., Fazekas, B. S., Woods, D., & Currow, D. C. (2005). The Australia-modified Karnofsky Performance Status (AKPS) scale: a revised scale for contemporary palliative care clinical practice [ISRCTN81117481]. BMC palliative care, 4, 7. doi:10.1186/1472-684X-4-7

Collet, J.P. (2000). Limite des essais cliniques. La Revue du Praticien 50(8):833-837. French.

Kalbfleisch & Prentice (1980). Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley & Sons; 1980.

MathWorks (2019). Cox Proportional Hazards Model. Retrieved from https://www.mathworks.com/help/stats/cox-proportional-hazard-regression.html on 25 April 2019.

Mogensen, UB, Ishwaran, H, Gerds, TA. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software* 2012; 50(11): 1– 23.

Stefan Dietrich, Anna Floegel, Martina Troll, Tilman Kühn, Wolfgang Rathmann, Anette Peters, Disorn Sookthai, Martin von Bergen, Rudolf Kaaks, Jerzy Adamski, Cornelia Prehn, Heiner Boeing, Matthias B Schulze, Thomas Illig, Tobias Pischon, Sven Knüppel, Rui

Wang-Sattler, Dagmar Drogan, Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis, *International Journal of Epidemiology*, Volume 45, Issue 5, October 2016, Pages 1406–1420,

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24.

St Julien JB, Pinkerman R, Aldrich MC, et al: Poor survival for veterans with pathologic stage I non-small-cell lung cancer. Am J Surg 204:637-642, 2012 Crossref, Medline, Google Scholar

Appendix A

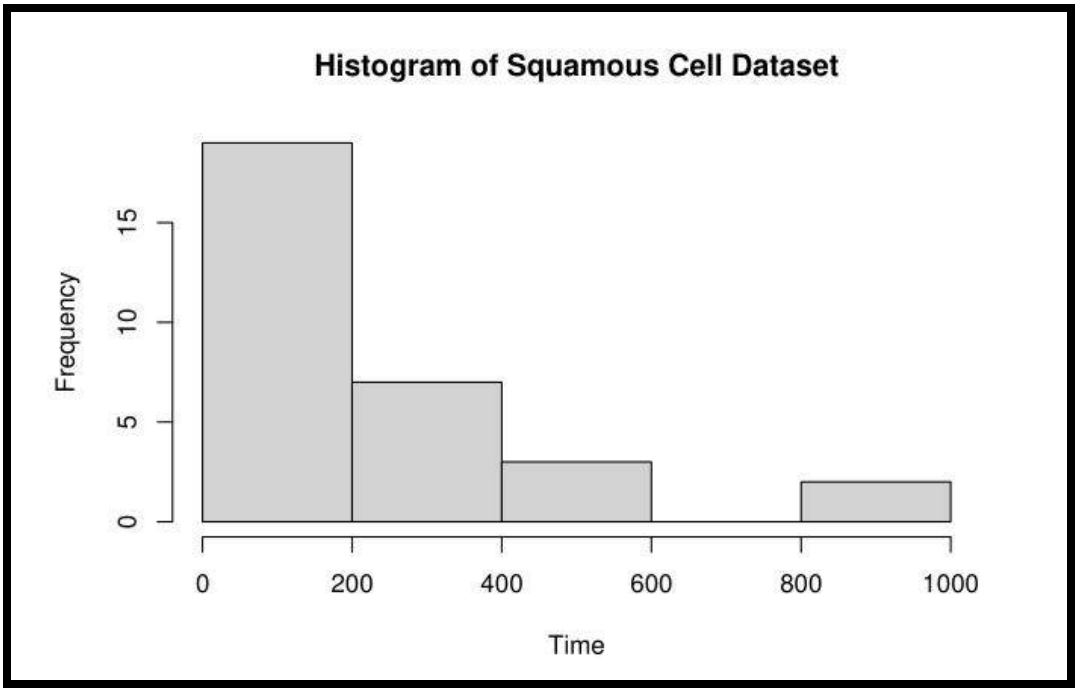


Figure 19: Histogram depicting survival time for the squamous cell subsection of the dataset

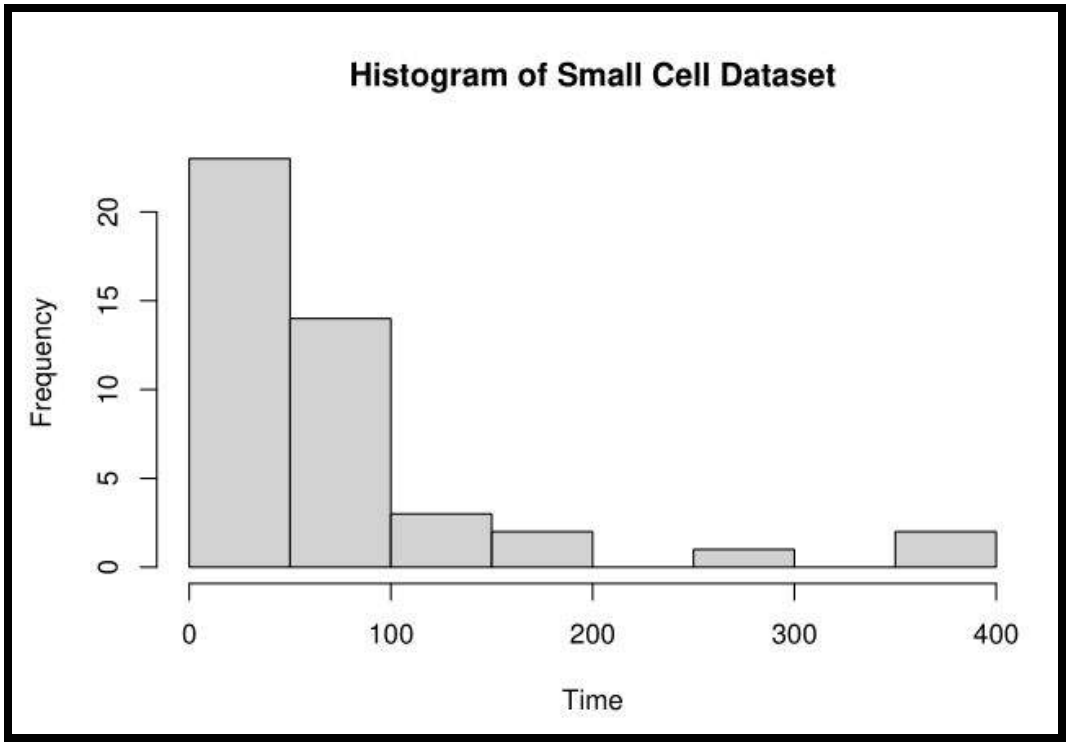


Figure 20: Histogram depicting survival time for the small cell subsection of the dataset

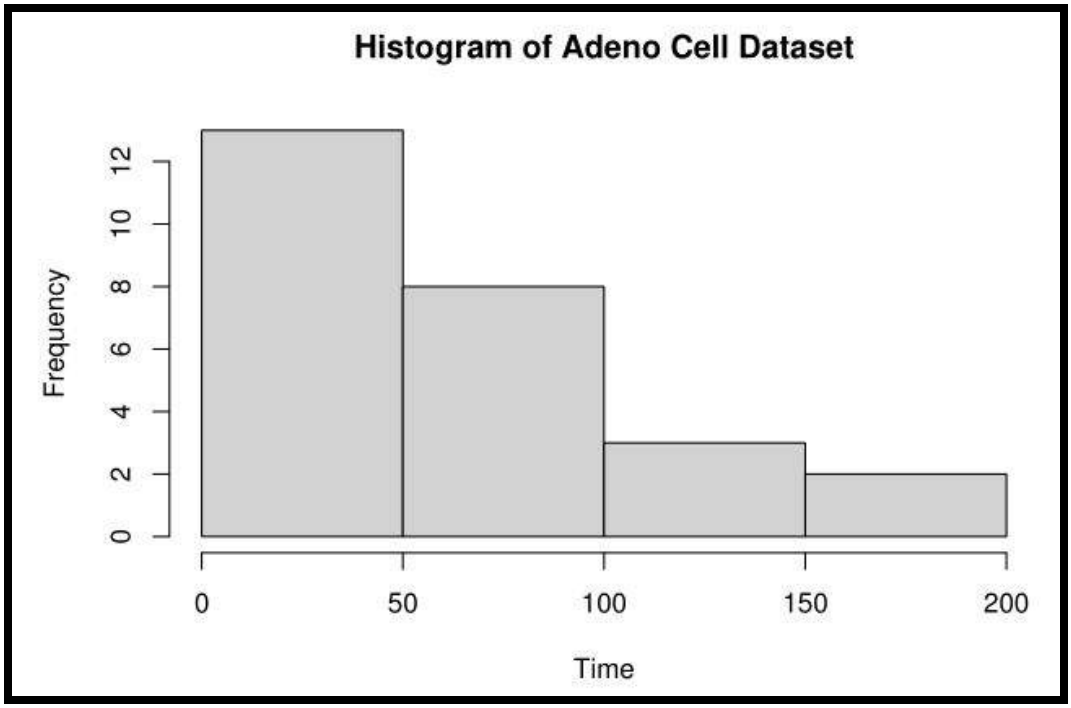


Figure 21: Histogram depicting survival time for the adenocarcinoma cell subsection of the dataset

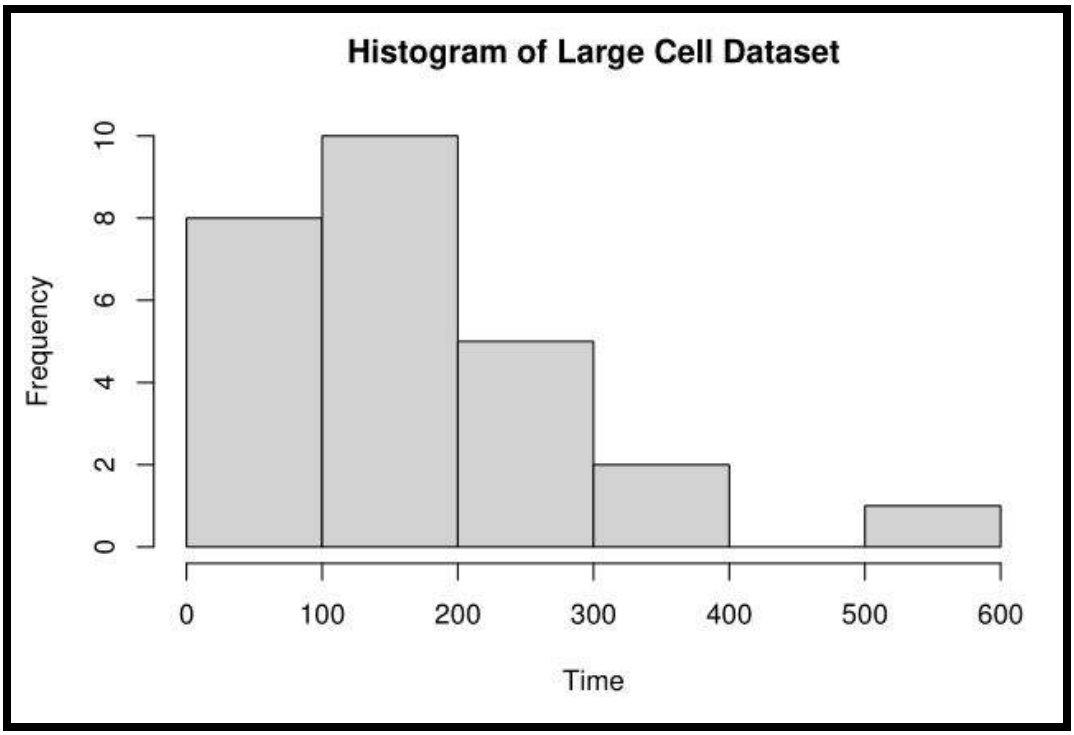


Figure 22: Histogram depicting survival time for the large cell subsection of the dataset

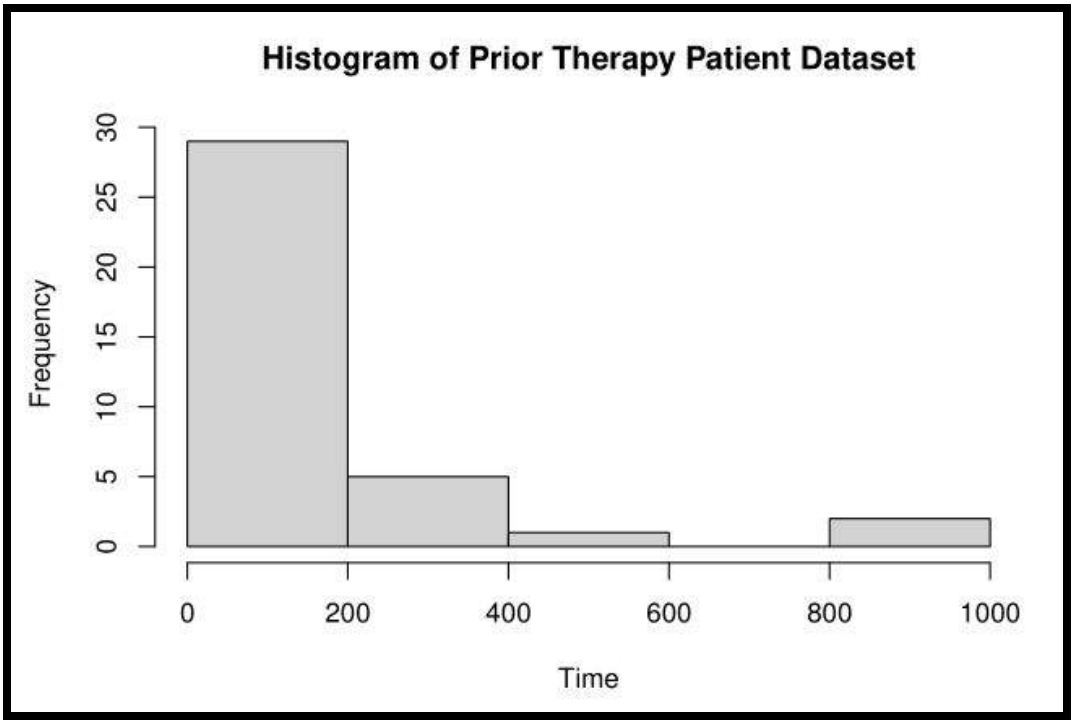


Figure 23: Histogram depicting survival time for the prior therapy subsection of the dataset

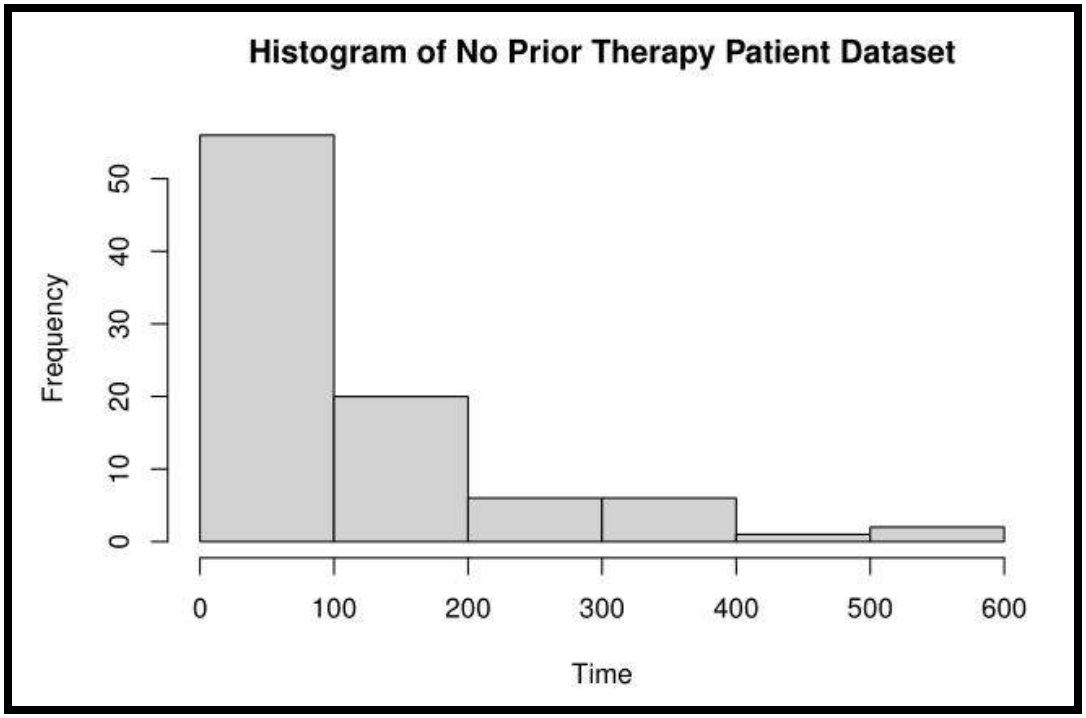


Figure 24: Histogram depicting survival time for the no prior therapy subsection of the dataset

Appendix B

R-Markdown File for Data Exploration and CPH Model

Please see attached file, “IEE573_RFiles_KMandCPH.pdf”

Jupyter Notebook for development of Cox Multilayer Perceptron for Deep Survival Analysis and Personalized Treatment Recommendations

Please see attached file, “IEE573_CoxMLP.pdf”