International Journal of Research in Engineering, Science and Management
Volume 4, Issue 8, August 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Development of a Machine Learning Model for Envisaging Cirrhosis Mortality Outcomes

Qusai Onali[1*], Om Brahmbhatt[2], Trupal Chaudhary[3]
[1,2]*B. Eng. Student, Department of Information Technology, G. H. Patel College of Engineering & Technology, Anand, India*
[3]*B. Tech. Student, U & P.U. Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, Anand, India*

***Abstract*: Rapid, standardized, and reliable prognoses can help physicians assess and administer their patients' most desirable treatment plans. Machine learning models can aid the physician decision-making process and provide a less biased estimate of patient status, which is very useful in determining transplant eligibility. We created models using data from longitudinal studies mapping patient prognosis over ten years. These models served to answer two questions: (1) can survival time be precisely foretold using easily obtainable biological indicators (2) can we predict patient disease progression (status) using these indicators. We found that using features such as albumin, bilirubin, edema, and ascites, a stepwise fit model could significantly predict survival time with a correlation coefficient of 0.66. We also found that disease status could be predicted at a 76% accuracy using logistic regression, random forest, and support vector machine (SVM) methods. The random forest model performed best at predicting survival status, and our three classification models prioritized similar features as the linear regression and each other. These top features align with current prognostics, which use variations of bilirubin, alkaline phosphatase, stage, and albumin in their predictive models, therefore supporting our initial hypothesis. The fact that there is only a single point of data in one study limits it. More preliminary testing and data collection should be done for future directions so that this model can be used clinically.**

***Keywords*: Python, Machine Learning, Cirrhosis.**

1. Introduction

Primary biliary cholangitis, earlier known as primary biliary cirrhosis, is a chronic disease in which the bile ducts in the liver are degraded [1]. It is a form of cirrhosis, a late stage of scarring in the liver caused by various liver diseases and disrupts liver function [2]. In 2018, roughly 4.5 million were diagnosed with liver disease, and roughly 41,700 people died from chronic liver diseases and cirrhosis in the United States alone [3].

We scrutinized a dataset from a research article published in 1989 titled "Prognosis in primary biliary cirrhosis: Model for decision making" to forecast the likelihood of survival for patients with primary biliary cirrhosis measurements that could be obtained through inexpensive, non-invasive methods.4 The authors began collecting data on patients at the start of the study in 1974 and continued collecting data from new patients for ten years. The data was collected on each patient at their appearance into the study, and their time of survival was updated throughout the study. At the time of the study, clinicians relied upon invasive liver biopsies to develop an accurate prognosis for these patients. Consequently, the researchers sought to develop a regression model based on non-invasive parameters that could provide an accurate prognosis and help clinicians determine whether the patient should receive a liver transplant. This model, which was built on many observations, could help clinicians generate a more standardized, accurate prognosis that they would provide based on their own experience.

The authors of the study developed their model using stepwise regression. They started with a set of 45 features and tapered down their model to 5 critical features, that is bilirubin (log), albumin (log), age, prothrombin time (log), and edema (and therapy) to predict a risk score that could be used to determine the probability of survival after *t* years. The dataset that was available online contained 312 observations by 19 features (Table A1). Each patient is represented by one observation.

It was unclear what the original authors used for their dependent or response variable when running their stepwise regression to develop their model. We believe that they used risk scores previously obtained or calculated from an existing model, but we did not have access to that data, so we decided to use an alternative dependent variable for building our models.

Our objective was to develop various models to predict the survival time of cirrhosis patients and the survival status - whether a patient survived or not throughout the study. The survival time is the time from the start of the study to when the patient either died, received a liver transplant, or the study ended and is in units of days. The survival status is labeled as 0 for censored, 1 for transplant, and 2 for death, so we grouped censored and transplants together into the survival group.

Additionally, we sought to determine what features are most important in determining the prognosis of primary biliary cholangitis for a given patient. Today the increased bilirubin

[*] Corresponding author: trupal.v.chaudhary@gmail.com

and alkaline phosphatase levels are associated with worse outcomes. Also, cirrhosis, indicated by histologic stage 4, is associated with a worse prognosis. Furthermore, one model called the GLOBE score predictive model uses serum bilirubin, albumin, alkaline phosphatase, platelet count after one year of UDCA treatment, and age at the start of therapy. The UK-PBC score model includes serum alkaline phosphatase, aminotransferases, and bilirubin after 12 months of UDCA therapy, as well as baseline albumin and platelet count [7]. Therefore, we hypothesized that these changes mentioned above/indicators would more strongly correlate with a decreased survival time and occurrence of death.

2. Methods

We obtained the data from a GitHub repository and found documentation on the dataset from the R Documentation site [5], [6]. The GitHub file originally contained 418 observations with 19 features. Still, we took the first 312 observations as these contained data for all features (The additional observations with missing features corresponded to an independent test set used in the original study to validate their model). Therefore, the dataset we used for this paper, located in the file cirrhosis.csv, contains 312 observations by 19 features (Table A1).

We made separate models to predict survival time and survival status independently. We used various unsupervised learning techniques, namely principal component analysis (PCA) and k- means clustering, to explore any grouping patterns in our data for the survival status. Then, we created various regression and classification models using stepwise regression, linear regression, lasso, random forest, and a support vector machine to predict survival time and survival status, respectively. We use an alpha level of 0.05 to evaluate statistical significance regarding correlation coefficients, feature selection, and independent two-sample t-tests. This paper highlights the important findings and results from our data analysis. All of our code and remaining results can be found in the file finalProject.mlx. We also ran our entire analysis again after standardizing the data using z scores and compared these results to our initial results.

3. Results

A. Summary Statistics

We gathered data such as means, medians, standard deviations, and so on. We used histograms, box plots, scatterplots, and correlation matrices to see if there were any evident links between different features, notably for time and status. After missing values were removed, the data comprised 276 observations. We added three variables, log(albumin), log(bilirubin), and log (prothrombin time), because these were variables used by the researchers in their risk score model.

B. Initial Exploration Plots

We split each feature into two groups for the boxplots, low and high-risk, corresponding to high and low survivability, respectively. We calculated risk scores for every observation using the equation provided in the cirrhosis paper.[4] We used the median risk score, 4.68, as a cutoff for low and high risk. Figure 1L saw those variables such as age, albumin, bili, stage, status, and time were all different between low risk and high risk, and some of these differences matched what was expected. For example, the high-risk group in stage had a median value of 4, and bilirubin increased for the high-risk group, consistent with current prognostics. For the scatter plots (log(time) scatterplot seen in Figure 1R), we plotted each variable versus time and then again versus log(time). Most of these scatter plots did not yield any particular insights except for the following: it seemed that albumin had a linear relationship with time and low alkaline phosphatase (alkphos), low bilirubin, low cholesterol, and low prothrombin time were all highly clustered with larger log(time) values. This may suggest that when these features are lower, the patient has a longer survival time. We might expect to see some of these features selected when predicting time using stepwise regression, lasso, and random forest.

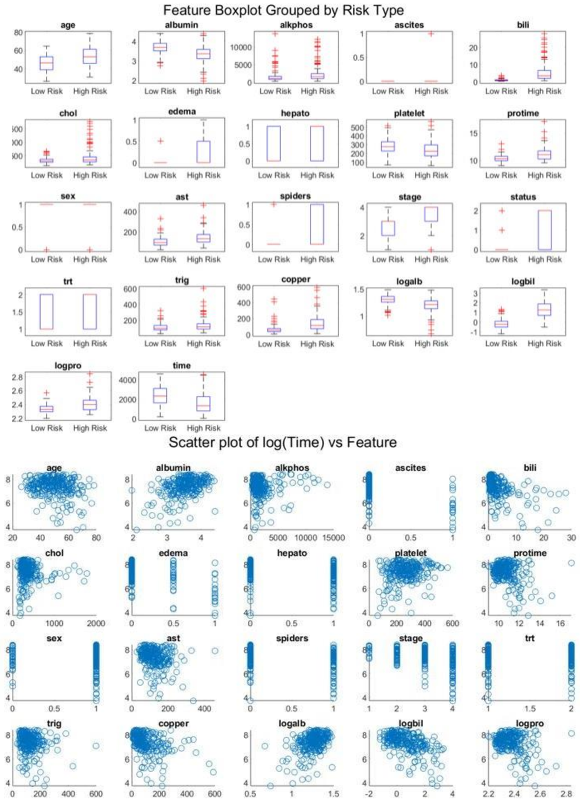


Fig. 1. On the top, T, are boxplots of the different features grouped by risk score. On the bottom, B, are scatter plots of the log(time) vs the corresponding feature

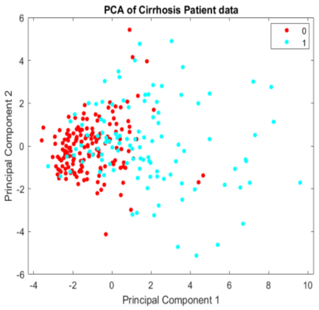
C. Statistical Analysis

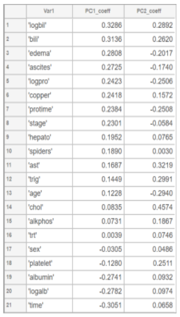
We calculated pairwise Spearman and Pearson's correlation coefficients for the data containing original time values and log(time) values. We discovered that the absolute correlation values between features and time were best for log(time) when looking at Pearson's coefficients. Thus, we opted to utilize log(time) for our supervised learning models' survival time response variable. Edema, log(bili), ascites, bilirubin, log(albumin), albumin, status, copper, stage, and log(albumin) were the top 10 features with the best absolute Pearson's coefficients for log(time) for Pearson's (from best to worst) (prothrombin time). Significant correlations of larger than 0.3 were found in all of them. Edema had a Pearson correlation of 0.5402 and a p-value of less than 1e-5. We later used these features to fit our linear regression model.

D. Principal Component Analysis

We ran PCA on the complete dataset after removing the status feature better to understand the correlations between all the variables and status and generated a scatter plot of the PC1 scores vs. PC2 scores, with status as the grouping variable. Each of the variables explains the percentages of the total variance.

PC1 had 27.07 percent, and PC2 had 11.17 percent, respectively. As shown in Figure 2L, PC1 successfully divided the data into two distinct groups. As a result, we sorted the features by PC1 coefficients and discovered that log(bilirubin), bilirubin, edema, ascites, and log(bilirubin) were the top five weighted features (prothrombin time). We used box plots to separate the groups based on survival status, and two sample independent t-tests were used to compare the two groups. They all showed significant differences (p 0.05), and many features grew in value between the afflicted and the survivors.





	Var1	PC1_coeff	PC2_coeff
1	'logbil'	0.3286	0.2892
2	'bili'	0.3136	0.2620
3	'edema'	0.2808	-0.2017
4	'ascites'	0.2725	-0.1740
5	'logpro'	0.2423	-0.2506
6	'copper'	0.2418	0.1572
7	'protime'	0.2384	-0.2508
8	'stage'	0.2301	-0.0584
9	'hepato'	0.1952	0.0765
10	'spiders'	0.1890	0.0030
11	'ast'	0.1687	0.3219
12	'trig'	0.1449	0.2991
13	'age'	0.1228	-0.2940
14	'chol'	0.0835	0.4574
15	'alkphos'	0.0731	0.1867
16	'trt'	0.0039	0.0746
17	'sex'	-0.0305	0.0486
18	'platelet'	-0.1280	0.2511
19	'albumin'	-0.2741	0.0932
20	'logalb'	-0.2782	0.0974
21	'time'	-0.3051	0.0658

Fig. 2. Left, PCA of cirrhosis data with status removed. The data is grouped nicely by PC1. Right, PCA contribution sorted by PC1

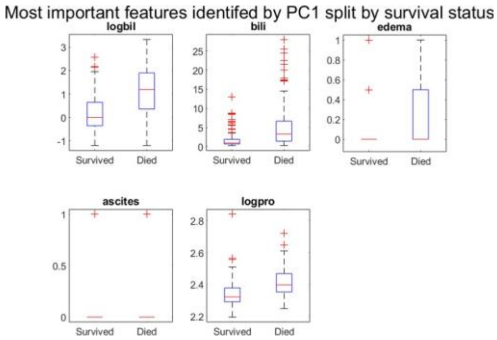


Fig. 3. Box plots of most important features identified by PC1 grouped by survival status. The two groups for each feature are significantly different as shown in the table below. For every feature except for ascites, the feature value increases for the group that did not survive

Table 1

Features	P values
'logbil'	7.43e-18
'bili'	3.58e-13
'edema'	1.54e-08
'ascites'	3.05e-07
'logpro'	3.81e-12

E. K-means

K-means was also applied to estimate cluster sizes of K=2:10. With a silhouette score of 0.9274, we learned that K=2 was the best cluster size. The silhouette plot, on the other hand, revealed that the features were divided quite disproportionately. We used the different cluster indexes to group the status variables and observed no discernible difference between the surviving and deceased groups.

Overall, we assumed that the stepwise regression, lasso, and random forest models would prioritize features that demonstrated some association with survival time and status, whether through local clustering, linear correlation, or PCA.

F. Supervised Learning

1) Predicting Survival Time

We developed six distinct models, each based on a different sample of data. The response variable in all of the models is log(time). Using the following data sets, we created three stepwise regression models:

1. Entire feature set
2. A subset of 12 noninvasive variables
3. Subset of 12 noninvasive variables with log values substituted for albumin, bili, & protime

We computed pairwise Pearson's correlation coefficients to choose ten features for linear regression, as earlier described in Background - Statistical analysis. We used the full feature set for lasso and random forest.

We used 5-fold cross-validation to verify each model. The average Pearson's coefficient, Spearman's coefficient, mean absolute error, and the number of coefficients (features) employed by the model were determined (Table A2, Figure 3). According to the average Pearson's correlation coefficient, lasso outperformed random forest.

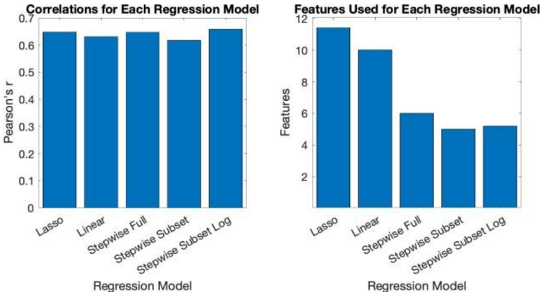


Fig. 3. Graphical summary of Pearson's R and the number of coefficients (features) used for each model

2) Predicting Survival Status

To predict survival status, we developed three models: a logistic regression model, a random forest classification model, and a support vector machine for binary classification. We eliminated the time feature from the data and ran 5-fold cross-validation once more. With an average Pearson's correlation coefficient of 0.54159 and an accuracy of 0.77896, we discovered that the random forest model performed the best. On the other hand, the logistic regression model fared poorly, with a Pearson's correlation coefficient of 0.47797 and an accuracy of 0.75013. (Table A3). Figure 4 shows the average classifications from all three models as confusion matrices.

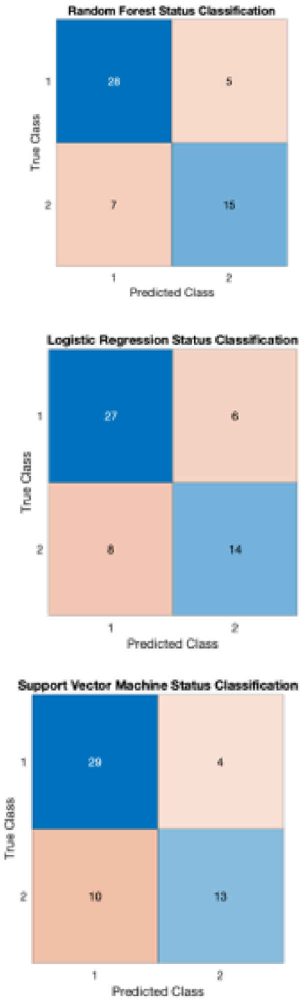


Fig. 4. Average Confusion Matrices for Random Forest and Logistic Regression

3) *Standardizing Data*

We used z scores to normalize our data and reran all of our unsupervised and supervised studies. We discovered that supervised models performed worse in general than unsupervised models. Thus, we present our preliminary findings here. A separate file called finalProjectZscore.mlx contains all of the z score data results.

4. Discussion

A. *Survival Time Models and Feature Selection*

Table A2 and Figure 3 show that the stepwise subset log performed best in predicting survival time. Among the models illustrated in Figure 3, it likewise had the fewest characteristics (5.2). The second stepwise regression model (Stepwise subset) had the smallest number of features (5.0) and the lowest Pearson's R (0.617), but it performed similarly to the other models. A model with fewer characteristics is preferable since it reduces overfitting and, from a clinical standpoint, requires fewer patient data to generate a prognosis, saving time and resources for physicians.

Albumin, bilirubin, edema, ascites, copper, stage, and alkaline phosphatase consistently surfaced among the top characteristics in all regression models except random forest (not necessarily in the order listed). (These findings may be found in Task 5 - Timed Supervised Learning.) This is consistent with current prognostics, which incorporate bilirubin, alkaline phosphatase, stage, and albumin changes into their models, confirming our initial idea. These characteristics also match the Pearson's correlation coefficients for log(time) in order (see Task 3 - Statistical Analysis).

B. *Survival Status Models and Feature Selection*

Overall, the models performed relatively similarly; however, based on Table A3, the random forest model did the best, even though this is not statistically significant. Similar traits were ranked as having the most value in all three classification models. Bilirubin, prothrombin time (and log (prothrombin time)), albumin, copper, and alkaline phosphatase obtained the greatest significance values in a random forest, according to the out-of-bag feature importance graphs. This was identical to the top five features chosen by the logistic regression model, which included copper, alkaline phosphatase, and log (prothrombin time). In contrast to logistic regression, which prioritized age and aspartate aminotransferase in separate validation runs, logistic regression prioritized age and aspartate aminotransferase consistently. Bilirubin and log (prothrombin time) were among the top five features chosen by PCA from among the above features. These characteristics are in line with current physiological indicators and prognostic models. Bilirubin, for example, appears to be particularly essential in predicting survival status, which is consistent with what has been shown in recent laboratory research studies. According to one report, bilirubin levels rise as the disease advances and are notably high in patients who have evident clinical symptoms.[8]

5. Conclusions and Future Scope

The study equaled a ten-year longitudinal study, but there is only one data point per patient in the dataset. More data points could have strengthened our model’s ability to add risk scores and disease status with this simplified data analysis. Taking measurements more consistently throughout the study could have made for a more robust model capable of predicting changes over time. The other limitation is the time variable itself. It is defined as “days between registration and earliest of death, liver transplantation and July 1986,” this means that all patients that survived past July 1986 were assumed to have died on that date, which could have negatively affected our model’s predictive abilities.

We ran a wide range of tests on this dataset in terms of data analysis. The models are severely limited by the data supplied, which is one of the most important takeaways. All of the models we put to the test functioned admirably. Rather than testing all of the different algorithms to see which one performs best, the model should be chosen based on the prediction and the available explanatory variables. Data that is well-formatted and includes many observations is crucial in developing powerful predictive models for research like these. We had a hard time locating biological datasets that would allow us to build the machine learning models we mentioned. Future directions of this project would be to gather more data and train/test these models on larger populations to see how they fare.

References

[1] Primary biliary cholangitis. (2018, March 09). Retrieved April 14, 2020. https://www.mayoclinic.org/diseases-conditions/primary-biliary-cholangitis-pbc/symptoms-causes/syc-20376874

[2] Cirrhosis. (2018, December 07). Retrieved April 13, 2020. https://www.mayoclinic.org/diseases-conditions/cirrhosis/symptoms-causes/syc-20351487

[3] FastStats - chronic liver disease or Cirrhosis. (2013, May 30). Retrieved April 13, 2020, https://www.cdc.gov/nchs/fastats/liver-disease.htm

[4] Dickson, E.R., Grambsch, P.M., Fleming, T.R., Fisher, L.D. and Langworthy, A. (1989), "Prognosis in primary biliary cirrhosis: Model for decision making," Hepatology, 10: 1-7.

[5] Therneau, T. (2017, March 27). Therneau/survival. Retrieved April 14, 2020, https://github.com/therneau/survival/blob/master/data/pbc.rda

[6] Pbcseq. (n.d.). Retrieved April 14, 2020. https://www.rdocumentation.org/packages/survival/versions/3.1-11/topics/pbcseq

[7] Clinical manifestations, diagnosis, and prognosis of primary biliary cholangitis (primary biliary cirrhosis). (n.d.). Retrieved April 14, 2020. https://www.uptodate.com/contents/clinical-manifestations-diagnosis-and-prognosis-of-primary-biliary-cholangitis-primary-biliary-cirrhosis#H17

[8] Reshetnyak, "Primary biliary cirrhosis: Clinical and laboratory criteria for its diagnosis." World J Gastroenterol. 2015;21(25):7683–7708.

Appendix

Table A1
A summary of the variables in the raw data with brief descriptions, units, or binary classification for categorical variables, and min, median, and max values

Var	Description	Code	Min	Median	Max
age	age	years	26.278	49.71	78.439
albumin	albumin	gm/dl	1.96	3.545	4.4
alkphos	alkaline phosphatase	U/liter	289	1277.5	13862
ascites	ascites	0 = no 1 = yes	0	0	1
bili	serum bilirubin	mg/dl	0.3	1.4	28
chol	serum cholesterol	mg/dl	120	310	1775
edema	edema treatment	0 = no edema 0.5 = untreated or successfully treated 1 = edema despite diuretic therapy	0	0	1
hepato	hepatomegaly	0 = no 1 = yes	0	1	1
time	time	days between registrationand earliest of death, liver transplantation and July 1986	41	1788	4556
platelet	platelets	count per mm^3 blood/1000	62	257	563
protime	prothrombin time	seconds	9	10.6	17.1
sex	sex	0 = male 1 = female	0	1	1
ast	aspartate aminotransferase, once called SGOT	U/ml	28.38	116.62	457.25
spiders	spiders	0 = no 1 = yes	0	0	1
stage	stage	1,2,3,4	1	3	4
status	censoring	0 = censored 1 = transplant 2 = death	0	0	1
trt	treatment	1 = D-penicillamine 2 = placebo	1	2	2
trig	triglycerides	mg/dl	33	108	598
copper	urine copper	micrograms/day	4	74	588

Table A2
Overview of the supervised learning models by averaging the cross-validation results

Evaluation Metrics (Average)	Stepwise Full	StepwiseSubset	Stepwise Subset Log	Linear Fit Lm	Lasso	RandomForest
Pearson's Correlation	0.6467	0.617	0.6588	0.6305	0.647	0.6488
Rank Correlation	0.5632	0.5578	0.5627	0.5259	0.5549	0.5747
Mean Absolute Error	0.4638	0.4835	0.4637	0.4744	0.4685	0.4551
Number of Coefficients	6	5	5.2	10	11.4	NA

Table A3
Logistic Regression versus Random Forest for predicting status

	Logistic	Random Forest	Support Vector Machine
Pearson's Correlation	0.47797	0.54159	0.4832
Mean Absolute Error	0.24987	0.22104	0.24273
Accuracy	0.75013	0.77896	0.75727