David Chang, Mahmoud Komaiha
BME 499 060: AI in BME

Creating a Machine Learning Model to Predict Cirrhosis Survival Outcomes

Abstract

Fast, standardized and reliable prognoses can help physicians evaluate and provide the best treatment plans for their patients. Machine learning models can support the physician decision making process and provide a less biased estimate of patient status, which is especially useful in determining transplant eligibility. We created models using data from longitudinal studies mapping patient prognosis over a 10 year period. These models served to answer two questions: (1) can survival time be accurately predicted 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. One study is limited by the fact that there is only a one-time point of data. For future direction, more preliminary testing should be done and data collected so that this model can be used clinically.

Introduction

Primary biliary cholangitis, previously known as primary biliary cirrhosis, is a chronic disease in which the bile ducts in the liver are slowly degraded.[1] It's a form of cirrhosis, which is a late stage of scarring in the liver that is 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 examined a dataset from a research article published in 1989 titled “Prognosis in primary biliary cirrhosis: Model for decision making” that aimed to predict the likelihood of survival for patients with primary biliary cirrhosis based on measurements that could be obtained through inexpensive, noninvasive methods.[4] The authors began collecting data on patients at the start of the study in 1974 and continued collecting data from new patients over the course of the 10 years. The data was collected on each patient at their entrance into the study, and their time of survival was updated over the course of the study. At the time of the study, to develop an accurate prognosis for these patients, clinicians relied upon invasive liver biopsies. Therefore, the researchers sought to develop a regression model based on noninvasive 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 than they would be able to 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 narrowed 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 to us 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 they had been previously obtained or calculated from an existing model, but we did not have access to that data, so we decided that we would 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 over the course of 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 transplant together into the survival group.

Furthermore, we sought to determine what features are most important in determining the prognosis of primary biliary cholangitis for a given patient. It's known today that increased bilirubin and alkaline phosphatase levels are associated with worse outcomes. Also, the presence of cirrhosis, indicated by the 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 aforementioned changes/indicators would more strongly correlate with a decreased survival time and occurrence of death.

Methods

We obtained the data from a Github repository and found documentation on the dataset from the RDocumentation site.[5,6] The Github file originally contained 418 observations with 19 features, but 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, which is 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 with respect to 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 for statistical significance with regards to 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.

Results

Summary Statistics

We obtained means, medians, standard deviations, etc. and made histograms, box plots, scatterplots and correlation matrices to observe whether there were obvious relationships between different features, especially with respect to time and status. The data had 276 observations after removing missing values. 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.

Initial Exploration Plots

For the boxplots, we split each feature into two groups, 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. From Figure 1L, we saw that 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, which was 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 for when predicting for time using stepwise regression, lasso and random forest.

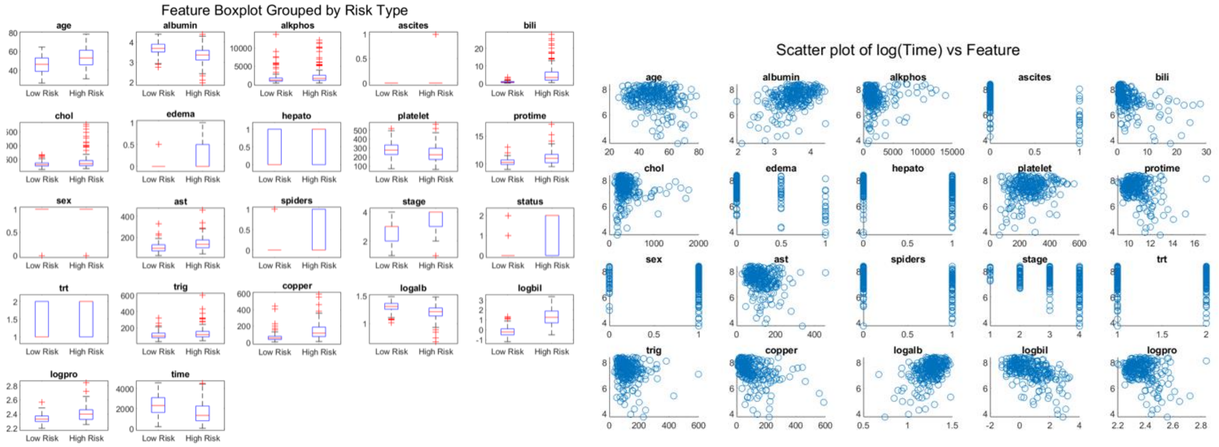


Figure 1. On the left, L, are boxplots of the different features grouped by risk score. On the right, R, are scatter plots of the log(time) vs the corresponding feature.

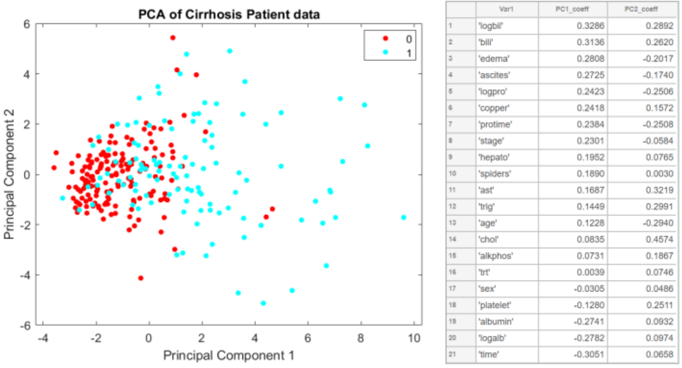
Statistical Analysis

We calculated pairwise Spearman and Pearson's correlation coefficients for the data with original time values and with log(time) values. When looking at Pearson's coefficients, we found that the absolute correlation values between the features and time were best for the log(time), so we decided to use log(time) for the survival time response variable for our supervised learning models. The top 10 features with the best absolute Pearson's coefficients with respect to log(time) for Pearson's were the following (from best to worst): edema, log(bili), ascites, bilirubin, log(albumin), albumin, status, copper, stage, and log(prothrombin time). They all had significant correlations greater than 0.3. 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.

Principal Component Analysis

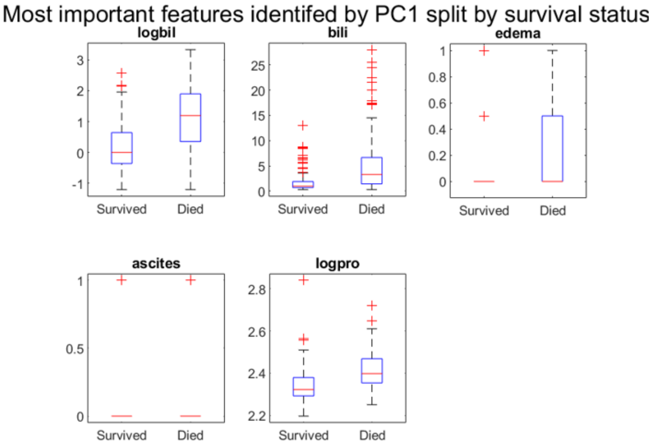
To better understand the relationships between all the features and status, we ran PCA on the entire dataset after removing the status feature and made a scatter plot of the PC1 scores vs PC2 scores, with status as the grouping variable. The percentages of the total variance explained by

PC1 and PC2 were 27.07% and 11.17%, respectively. PC1 grouped the data well into two distinct groups, as seen in Figure 2L. Because of this, we sorted the features by PC1 coefficients and found that the top five weighted features were log(bilirubin), bilirubin, edema, ascites, and log(prothrombin time). We made box plots for these features, separating the groups by survival status, and two sample independent t-tests were conducted between the two groups. They were all shown to have significant differences ($p < 0.05$) and many of the features increased in value between the survived and diseased groups.



	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

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



Top 5 PC1 Features

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

Figure 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 on the right. For every feature except for ascites, the feature value increases for the group that did not survive.

K-means

We also ran K-means and tested cluster sizes of K=2:10. We found that K=2 was the optimal cluster size with a silhouette score of 0.9274, however, the silhouette plot showed that the features were split very disproportionately. We grouped the status values by the different cluster indexes and found that there was no meaningful separation between the survived and deceased groups.

Overall, we hypothesized the features that showed some relationship with survival time and status, whether that be local clustering, linear correlation, or through PCA, would be selected for as having greater importance by the stepwise regression, lasso, and random forest models.

Supervised Learning

Predicting Survival Time

We created 6 unique models, each one using a different subset of data. All of the models use log(time) as the response variable. We made 3 stepwise regression models using the following sets of data:

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

For linear regression, as previously mentioned in Background - Statistical analysis, we calculated pairwise Pearson's correlation coefficients to select 10 features. For lasso and random forest, we used the entire feature set.

To validate each model, we performed 5-fold cross-validation. To evaluate model performance, we calculated the average Pearson's coefficient, Spearman's coefficient, mean absolute error, and the number of coefficients (features) used by the model (Table 1, Figure 3). Based on the average Pearson's correlation coefficient, lasso performed best, followed by random forest.

Evaluation Metrics (Average)	Stepwise Full	Stepwise Subset	Stepwise Subset Log	Linear FitLm	Lasso	Random Forest
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 1. Overview of the supervised learning models by averaging the cross-validation results.

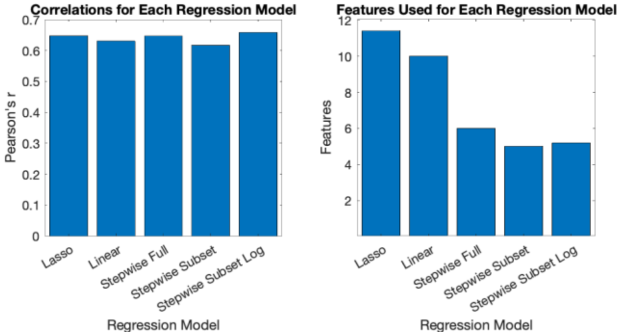


Figure 3. Graphical summary of Pearson's R and the number of coefficients (features) used for each model.

Predicting Survival Status

We created three models, a logistic regression model, a random forest classification model, and a support vector machine for binary classification to predict survival status. We removed the time feature from the data and once again performed 5-fold cross-validation. We found that overall, the random forest model performed best with an average Pearson's correlation coefficient of 0.54159 and an accuracy of 0.77896. This is in contrast to the logistic regression model, which performed the worst with an average Pearson's correlation coefficient of 0.47797 and an accuracy of 0.75013 (Table 2). Figure 4 contains confusion matrices showing the average classifications from all three models.

	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

Table 2. Logistic Regression versus Random Forest for predicting status.

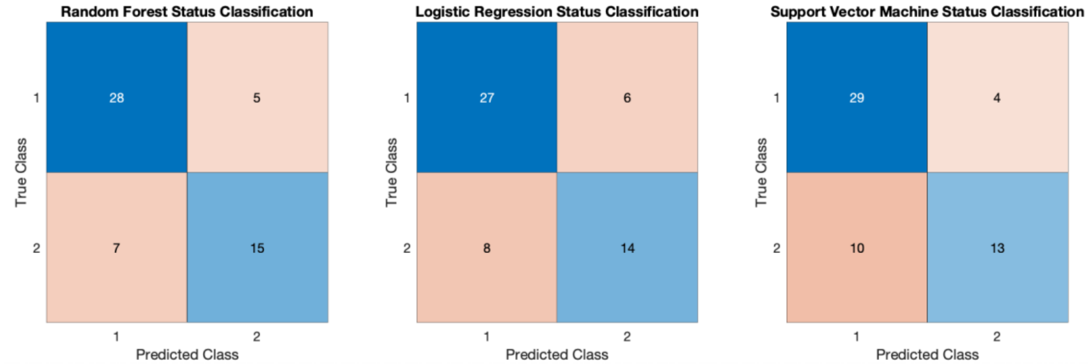


Figure 4. Average Confusion Matrices for Random Forest and Logistic Regression.

Standardizing Data

We standardized all our data using z scores and reran all of our unsupervised and supervised analyses. We found that the supervised models overall performed worse than with the data not standardized, so we only report on our initial results here. All the results with the z score data can be found in a separate file called finalProjectZscore.mlx.

Discussion

Survival Time Models and Feature Selection

For predicting survival time, stepwise subset log performed best, as shown in Table 1 and Figure 3. It also used the least number of features on average (5.2) amongst the models shown in Figure 3. The second stepwise regression model (Stepwise subset) used the least number of features (5.0) and had the lowest Pearson's R of 0.617, but this was still close to how the other models performed. Having a model with fewer features is favorable not only because it helps to prevent overfitting, but also from a clinical standpoint, less patient data is required to generate a prognosis, saving clinicians time and resources.

Across all regression models excluding random forest, albumin, bilirubin, edema, ascites, copper, stage, and alkaline phosphatase consistently appeared within the top features (not necessarily in the order listed). (These results can be seen in Task 5 - Supervised Learning on Time). This aligns with current prognostics, which use variations of bilirubin, alkaline phosphatase, stage, and albumin in their predictive models, therefore supporting our initial hypothesis. These features also align with the order of Pearson's correlation coefficients with respect to log(time) (see Task 3 - Statistical Analysis). For more details on which features were selected during each cross-validation round for each model, please see our live script.

Survival Status Models and Feature Selection

Overall the models performed very similarly, while it is not statistically significant, the random forest model performed the best based on Table 2. The three classification models prioritized similar features as having the greatest importance. For random forest, the out-of-bag feature importance graphs show that bilirubin, prothrombin time (and log(prothrombin time)), albumin, copper, and alkaline phosphatase had the highest importance values. This was similar to the features selected by the logistic regression model, which also selected copper, alkaline phosphatase, and log(prothrombin time) within the top 5 features. Logistic regression differed in that it prioritized age and aspartate aminotransferase consistently across different validation runs. Out of the features mentioned above, bilirubin and log(prothrombin time) were included in the top

5 features selected by PCA. These features align with current predictive models and physiological indicators. For example, Bilirubin, seems to be extremely important in predicted survival status, which supports what has been found in modern laboratory research studies. According to one article, bilirubin levels elevate as the disease progresses and are especially elevated in patients that show obvious clinical symptoms.[8]

Conclusions and Future Directions

This study was a 10 year longitudinal study, but in the dataset used there is only one data point per patient. While this simplified data analysis more data points could have strengthened our models ability to interpolate risk scores and disease status. By taking measurements more consistently over the duration of 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.

In terms of data analysis, we performed a large variety of tests on this dataset. One of the biggest take-a-ways is the models are very severely limited by the data available. All of the models we tested performed well. Instead of taking the approach of trying all of the different algorithms and seeing what works best, the model should be chosen based on what is being predicted and what explanatory variables are available. For studies like these, data that is well-formatted and has a large number of observations is critical in making strong predictive models. We had difficulty finding biological datasets that were conducive to developing the machine learning models discussed. Future directions of this project would be to gather more data and train/test these models on larger populations to see how they fare.

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 registration and 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

References

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

[2] Cirrhosis. (2018, December 07). Retrieved April 13, 2020, from 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, from 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. doi:10.1002/hep.1840100102

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

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

[8] Reshetnyak VI. Primary biliary cirrhosis: Clinical and laboratory criteria for its diagnosis. World J Gastroenterol. 2015;21(25):7683–7708. doi:10.3748/wjg.v21.i25.7683