Analysis of Mortality Following Heart Failure

Naive Frequentists (Group 3)
Addison McGhee
Zach Clement
Dan Nolte

I. Abstract

Heart failure is a serious chronic condition that affects over half a million Americans per year. Although many people live with heart failure for years, for others death follows soon after diagnosis. It is of great clinical importance to evaluate a patient's risk of death following heart failure, so that clinicians can appropriately treat and communicate risks with patients. This study uses a dataset of 299 heart failure patients and their health characteristics and survival status to build a logistic regression model which predicts 30-day mortality following heart failure, and assesses which health characteristics are most predictive of death for these patients. Separately, this study uses survival analysis to analyze the expected remaining lifespan of a patient given their health characteristics presenting during heart failure diagnosis. Our prediction model achieved an accuracy of 0.66 at the threshold which best balanced sensitivity and specificity. We found that age, serum creatinine, and serum sodium were the most predictive characteristics of death. A survival analysis selected these three variables as well as an interaction between serum creatinine and serum sodium, ejection fraction, and quadratic ejection fraction.

I. Introduction

The term "heart failure" refers to a condition where the heart is not pumping blood as well as it should be. It is a chronic condition that over half a million Americans are diagnosed with each year.[1] Heart failure most often occurs in adults, and many people who develop heart failure have had often another heart condition previously

[1]Emory Healthcare, "Heart Failure Statistics", https://www.emoryhealthcare.org/heart-vascular/wellness/heart-failure-statistics.html

(e.g., coronary heart disease, high blood pressure, heart attack).[2] It is possible for people to live many years with heart failure with only mild quality of life implications. However, many people who have heart failure experience fatigue and shortness of breath, and can have difficulty walking or climbing stairs. Heart failure is a broad diagnosis, and any two people experiencing heart failure may have different underlying causes and prognoses. Physicians classify heart failure in several ways to guide treatment and assign risks to patients:

HFrEF vs. HFpEF: Heart failure typically occurs in the left side of the heart, and when it does it is classified as either "heart failure due to reduced ejection fraction" (HFrEF) or "heart failure with preserved ejection fraction" (HFpEF). Ejection fraction is a measure of the percentage of blood contained in the left ventricle that it pumps out during each heartbeat. HFrEF is when the left ventricle can't pump enough blood into circulation to meet the body's needs (i.e., the ejection fraction is too low, or below 40%). HFpEF, on the other hand, occurs when the left ventricle becomes stiff and cannot fill with enough blood during each rest in between each beat.[3] Our dataset contains both patients with HFrEF and HFpEF.

NYHA Classification: Physicians classify patients with heart failure on a scale called the NYHA Functional Classification which ranges from I-IV, where higher levels represent patients with more severe cases of heart failure who are more limited in their physical activities. Our dataset only includes patients classified as III or IV who have been diagnosed with severe heart failure.[4]

---

[2] Heart.org, "What is Heart Failure?", https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure
[3] Heart.org, "Types of Heart Failure", https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/types-of-heart-failure
[4] Heart.org, "Classes of Heart Failure", https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

Since heart failure has many risk factors and gradations of severity, the risk of death following heart failure is an oft studied question, and one we seek to ask as well. In particular, we would like to estimate the risk of death in the 30 days following a class III or IV heart failure diagnosis based on the characteristics that patients present at the time of diagnosis. Separately, we seek to assess the characteristics that are most associated with increased hazard of death following heart failure, and understand how each factor impacts survival time for each patient.

II. Review of Literature and Domain Expertise

There has been a good deal of research evaluating determinants of poor outcomes following heart failure. For example, in “Predicting Poor Outcomes in Heart Failure”, David Smith concluded that extreme hemoglobin levels and poor renal function are associated with mortality post-heart failure. This study also evaluated the impact of ejection fraction and left-ventricular wall thickness on death, and found both to be strong predictors of death. Like we intend to do, this paper focuses on identifying predictors of poor outcomes and does not concern itself with causal questions.[5]

In Feb. 2020, David Chicco published a paper which uses the same dataset that we intend to use. As a result, it is a very valuable resource, as we can compare our approaches to those of the authors, who happen to ask very similar questions to ours. One important learning from this paper is that their description of the dataset contains additional information not available at the source we obtained it. Most crucially, (1) they confirmed it is unknown how follow-up was carried-out, (2) they specified how anemia was defined, and most importantly, (3) they stated that “all 299 patients had left

[5] Smith DH, Johnson ES, Thorp ML, et al. Predicting poor outcomes in heart failure. Perm J. 2011;15(4):4-11.

ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure."[6] Since only those with severe heart failure are included in the study, we are less concerned about confounding by severity of "heart failure."

Chicco published another paper using this same dataset that we intend to use, but for the purpose of survival analysis. The authors of the paper find that there is a large risk of mortality in the days immediately following heart failure, but that this risk is gradually reduced for the remainder of the follow-up period. The authors found that age, renal dysfunction, blood pressure, ejection fraction and anemia were the greatest risk factors for death following heart failure.[7]

In 2013 Ana Alba published a meta-analysis of mortality prediction models like ours that have been published over the years. It attempts to externally validate these models on newer patient cohorts to evaluate their performance at predicting mortality among those with heart failure based on their characteristics. It finds inconsistent performance, and notes that the best models have "modest discrimination" and "questionable calibration," and postulates that poor performance has to do with evolving treatment and drug regimines which aren't captured by older prediction models.[8] It is important to acknowledge the inconsistency of prior prediction models for this task as we undertake our project, and to attempt to address the shortcomings of these past models that the authors highlight here.

---

[6] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020).
[7] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020).
[8] Alba AC, Agoritsas T, Jankowski M, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. Circ Heart Fail. 2013;6(5):881-889.

Indeed, our subject matter expert Nona Jiang, a cardiologist, also pointed out that it would be very valuable to know which medications each patient was taking in follow-up. She pointed out that there is a standard cocktail of medications called Guideline Directed Medical Therapy (GDMT) that is often administered to patients with heart failure; and the fact that we do not have this data available may introduce a source of unmeasured confounding. She also pointed out that there is a measurement commonly taken for heart failure patients not available in our dataset called B-type natriuretic peptide, or BNP, which measures the "stretch" of the heart. Separately, Nona suggested that we try testing the following interaction effects which have a plausible biological basis: creatinine and sodium; CPK and platelets; and Smoking and diabetes. We incorporate all of this feedback in our methodology and results sections.

III. Research and Analysis Methods

A. Data and Software

As discussed above, this dataset was provided by the UCI machine learning repository[9]. Models were fit using the glmnet and survival R packages. Visualizations were generated using the ggplot2, survminer, ggpubr, pRoc, and caret R packages. Goodness of fit tests were conducted using the ResourceSelection, survival, and survminer R packages.

B. 30-Day Mortality Prediction: Logistic regression

We fit a logistic regression model to predict 30-day mortality (individuals who had a recorded death event before 30 days). Individuals who were lost-to-follow-up before 30 days were removed from the analysis, and a complete case analysis was conducted.

[9] Chicco, D. (2020, January). UCI Machine Learning Repository: Heart failure clinical records Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

We believe this choice was justified because only 5 patients were lost to follow up before 30 days, which comprised only 1.7 percent of our full dataset.

In order to conduct this analysis, we assumed that the outcome mortality status followed a binomial distribution, with each patient's probability of 30-day survival being independent of other patients' survival. We also assumed that expected 30-day mortality status given our covariates was linear on the logit scale. An analysis of residual plots supported this assumption as none of the residuals indicated a non-linear trend. Lastly, we assumed that there were enough (10 to 20) events per covariate to use the generalized linear model framework, which was verified by the presence of 35 cases of 30-day mortality for three variables (11.667 events per covariate).

We then fit an elastic net regression model to perform variable selection and to produce estimates. We estimated tuning parameters (lambda and alpha) using a 5-fold cross-validation using the R glmnet package. Alpha values of 0, 0.1, 0.2, … 0.9, 1 were tested, and for each alpha value, 100 lambda values were tested. A final model was fit using the values of alpha and lambda which best minimized deviance. In addition, a logistic regression model using the variables selected by the elastic net model but without a penalty was fit.

Model fit was visually assessed using residual plots (see Supplemental Materials, **Figure 1**), a histogram of fitted probability values, and a plot of hat values against studentized residuals. Observations with high cook's distance and leverage were identified and a sensitivity analysis was conducted using a dataset without these observations (see Supplemental Materials, **Figure 2**). A Hosmer and Lemeshow goodness of fit test was conducted.

In order to evaluate model performance, predicted probability values were generated using 5-fold cross-validation. A cutoff which best balances sensitivity and specificity was chosen, and performance metrics were generated using that cutoff.

C. Survival Time Analysis

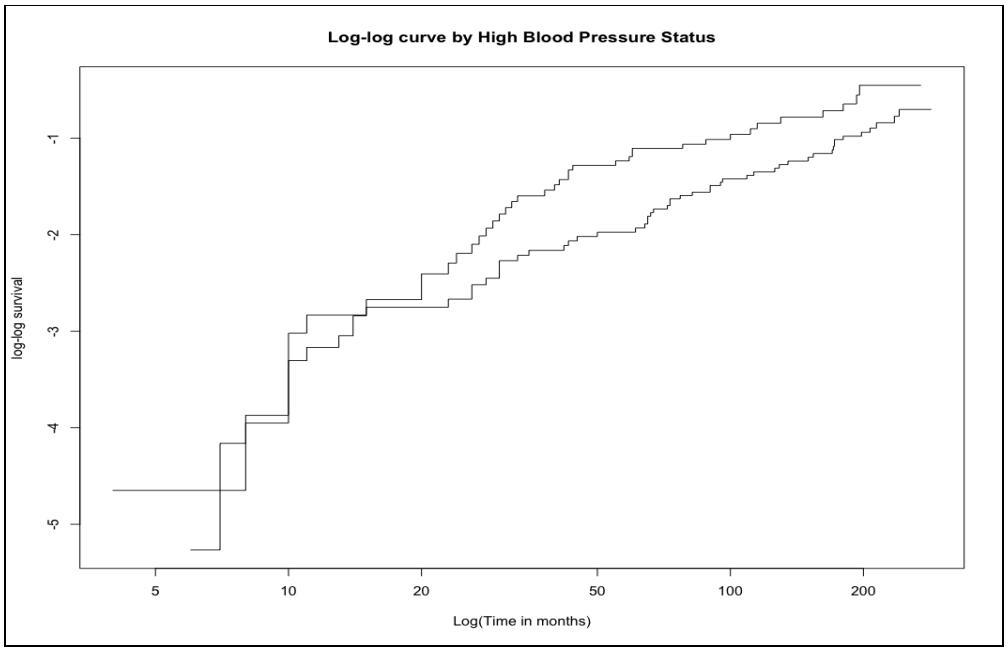
Our first step to analyze the Heart Failure dataset from a Survival Analysis perspective was to simply plot the Kaplan-Meier curves associated with this data, and generate median, 25th and 75th percentile survival times for patients presenting with heart failure. Continuing with non-parametric approaches, we then considered the binary variables in the dataset (amaemia, diabetes, high blood pressure, sex, and smoking); plotting survival curves for each to compare the two groups in each case. This allowed us to visually observe differences in survival across these strata, and helped inform variable selection and evaluation of the proportional hazards assumption later on when it came time to build a regression model.

We then considered a Cox Proportional Hazards Model. We began by assessing the suitability of this data for Cox regression by checking the assumptions of this model. Some assumptions (linearity, normality of betas, independence of observations) are infeasible to test and are otherwise reasonable to assume given what we know about this dataset. We then proceeded to fit a preliminary Cox model which could be used to evaluate the remaining assumptions.

In order to select variables for a first-pass model, we began by simply including all potential covariates in the model and reviewing which covariates were significant at the 0.05 level, which yielded 6 of the 11 covariates. Separately, we used the glmnet package in R to fit a Cox model with a LASSO penalty, selecting the lambda parameter

using 5-fold cross validation which resulted in the lowest average test-set MSE. This process sent all but 7 of the covariates to 0 (the same 6 as above, plus serum sodium). We then fit models using both sets of parameters and compared AIC values; ultimately selecting the model with 7 covariates as it had a lower AIC. We then considered the potential interaction effects delineated above and suggested by our subject matter expert; and ultimately decided to include the interaction of serum creatinine and serum sodium, which was significant and whose inclusion further lowered the model AIC. We then removed creatinine phosphokinase and anaemia to achieve a more parsimonious model, because neither had a biological significance according to our subject matter expert, and because doing so slightly decreased model BIC.

Before interpreting model results we first continued to ensure that model assumptions were met. As the final model had 7 covariates and the dataset had 94 death events, the 10 events per predictor assumption was met. We then checked the proportional hazards assumption by plotting, for each covariate, the Schoenfeld residuals and looking for non-zero slopes indicating time-variant beta values. We noticed that the trends of the Beta(t) for high blood pressure both showed a slight upward trend, especially at the right tail, so we further evaluated the PH assumption using the cox.zph test and log(-log(S)) vs. log(t) plots. We ultimately assessed that it did not violate the PH assumption and it remained in the model.



We then considered nonlinearities in the predictors. Our subject matter expert pointed out that an ejection fraction below 40% is severely unhealthy (and defines HFrEF), so we expected different hazards for patients above and below this threshold. We plotted Martingale residuals for this covariate and did in fact notice a curved relationship with an inflection point around 40 (see Supplemental Materials, **Figure 3**). We tried 1) coding HFrEF as a binary variable (HFrEF = 1 if ejection fraction was below 40), 2) fitting a cubic spline with a single knot at ejection fraction of 40, and including a quadratic ejection fraction term. The quadratic ejection fraction model performed best in terms of AIC/BIC, so we chose it as our final model.

Lastly we considered ties and influential points. There were ties in this dataset; in fact there are only 148 unique time points for 299 observations, implying that about two events (death or censorship) occurred at each time point on average. As a result, we fit our model using exact, Breslow, and Efron tie-handling methodologies and compared results. We found quite similar output in each case, suggesting low sensitivity

to tie-handling. In our final model, we used the default (Efron) tie-handling process. To evaluate high-influence points, we plotted dfbetas for each coefficient and didn't observe any observations exerting extremely high coefficient over the beta coefficients, and so didn't consider removing any observations from the dataset.

IV. Findings and Analysis

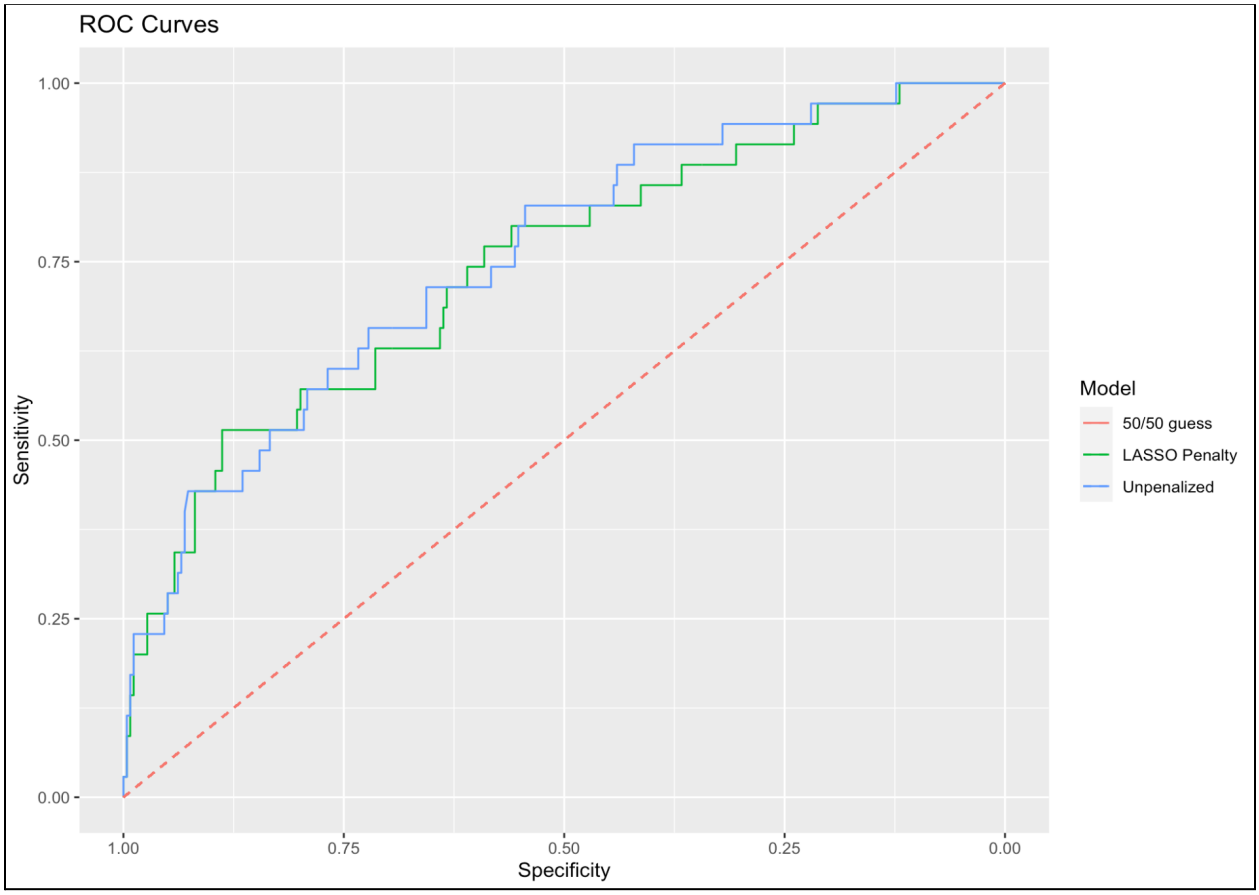
A. 30-Day Mortality Prediction

The values of α and λ which minimized deviance for a logistic regression model predicting 30-day mortality were 0 and 0.030, respectively. This is equivalent to fitting a LASSO model with λ = 0.030. Our fitted logistic model included an intercept as well as age, serum creatinine, and serum sodium as covariates to predict 30-day mortality. We used these selected covariates to fit a logistic regression model without the LASSO penalty, and we compared model performance for the two models. Model output is shown below.

	LASSO Estimate	Standard Estimate	Std. Error (standard)	z value (standard)	Pr(>|z|) (standard)
(Intercept)	0.1641068	5.2569536	5.4153570	0.9707492	0.3316732
age	0.0399350	0.0637748	0.0159922	3.9878745	0.0000667
serum_sodium	-0.0360385	-0.0869333	0.0391508	-2.2204745	0.0263866
serum_creatinine	0.1464361	0.2767879	0.1430254	1.9352361	0.0529613

Our logistic regression model with a LASSO penalty had a AUC of 0.7426 for the LASSO penalty model and 0.7539 for the unpenalized model. For the LASSO model, at the cutoff which best balanced sensitivity and specificity (0.1172), the model had accuracy 0.6395, specificity 0.64093, and sensitivity 0.62857. The unpenalized model had accuracy 0.6565, specificity 0.65637, and sensitivity 0.65714 at its cutoff which balanced sensitivity and specificity (0.102295). The ROC curve demonstrates that both

penalized and unpenalized models are better than a 50/50 guess at all levels of sensitivity and specificity.

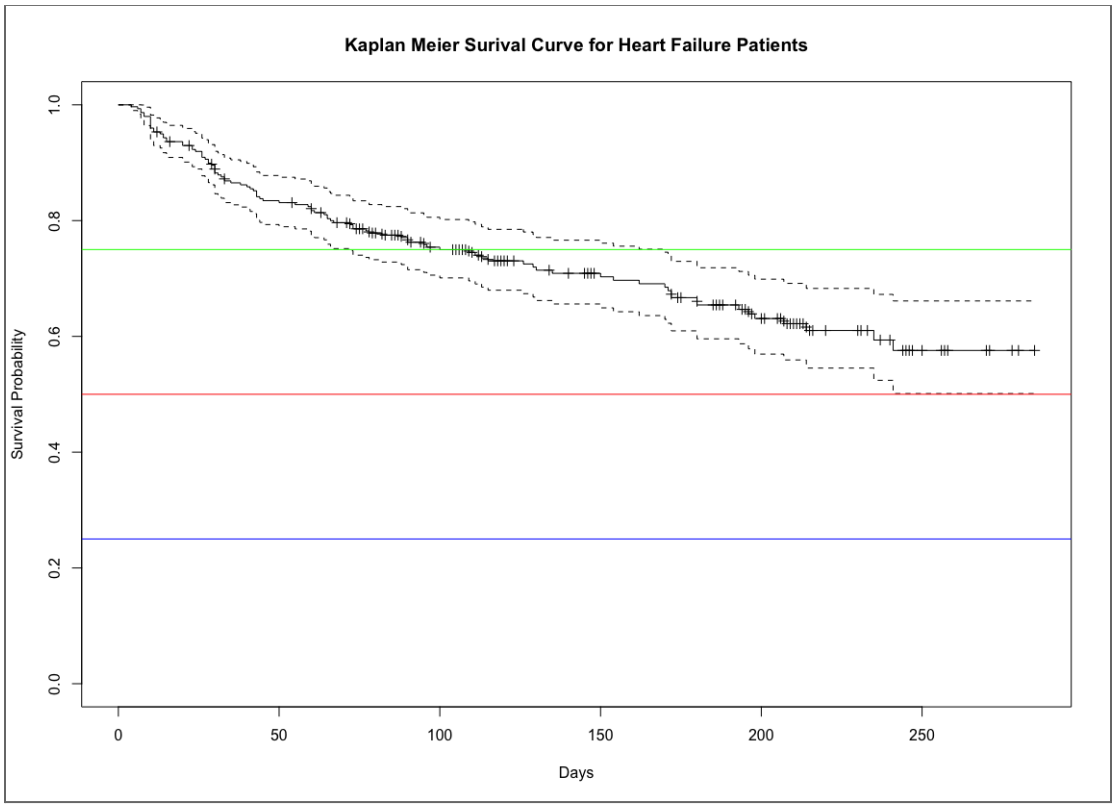


Because the logistic regression model without the LASSO penalty performed better, we will consider it as we assess goodness of fit. Residual plots of age, serum creatinine, and sodium all showed a fairly linear trend, suggesting that residuals were independent of predictors. A Hosmer and Lemeshow goodness of fit test returned a test statistic of 3.3539, with a p-value of 0.9102, which suggests that the model fit was adequate. The observations with the 3 highest hat values and the 3 highest cook's distance were identified. When a sensitivity analysis was conducted with these observations removed, the coefficient for age changed by less than 10%. The

coefficients for serum sodium and serum creatinine changed by more than 10%, but retained the same sign.

B. Survival Time Analysis

Below we present the output of the non-stratified Kaplan Meier curve associated with this dataset.



As we can see from the above figure, the 25th percentile survival time for heart failure patients is 100 days, and the median and 75th percentile survival times exceed 241 days (the longest follow-up period from this dataset.) The mean survival time, on the other hand, is only 130 days. This is because, as is apparent from the plot above, the risk of death is highest in the first 30 or so days following heart failure and levels off around 100 days. In other words, the large volume of deaths occurring directly following

heart failure greatly decreases the mean survival time to well below the median in this case.

The output of the final regression model, specified according to the description in the methods section above, is listed below:

	Coef	exp(Coef)	Std. Error	z value	Pr(>|z|)
age	0.0494150	1.0506563	0.0092099	5.365394	0.0000001
ejection_fraction	-0.1737614	0.8404974	0.0471845	-3.682591	0.0002309
ejection_fraction^2	0.0016236	1.0016249	0.0006030	2.692468	0.0070925
serum_creatinine	-3.1751814	0.0417865	2.5008048	-1.269664	0.2042044
serum_sodium	-0.0840170	0.9194156	0.0413301	-2.032827	0.0420700
high_blood_pressure	0.4074637	1.5030009	0.2150848	1.894433	0.0581676
serum_creatinine:serum_sodium	0.0252346	1.0255557	0.0184581	1.367131	0.1715842

It is clear from this model that age, ejection fraction, and serum sodium are the most significant predictors. From this output we may also conclude that the hazard of death following heart failure among those aged 65 is 1.63 [95% CI: 1.36-1.95] times the hazard of death of those aged 55 on average, holding all other covariates constant.

V. Discussion

A. Variable Selection

Age, serum sodium, and serum creatinine appear to be important for predicting mortality after heart attack. When a LASSO penalty was used, both a Cox proportional hazards model and a logistic regression model included age, serum sodium, and serum creatinine in the model. This is in concordance with suggestions by our subject matter expert that these variables may be related to heart disease. In addition, these three variables were selected by David Chicco[10] as most important for predicting heart

[10] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020).

disease. Our models add support to the theory that these variables are most helpful for predicting heart disease.

In both models, increasing age was associated with increased risk of mortality, and decreasing serum sodium was associated with increased risk of mortality. However, the direction of effect differed between logistic regression and Cox models for serum creatinine. This was likely due to the inclusion of other collinear covariates (such as the interaction between serum sodium and serum creatinine) in the Cox proportional hazards model.

The logistic regression model did not include four variables which were included by the Cox model: ejection fraction, quadratic ejection fraction, high blood pressure, and an interaction between Serum Creatinine and Serum Sodium. This may have been because a binary 30-day outcome carries less information than time-to-event data, so the model was not able to pick up on effects predicting longer-term mortality, meaning their effects on 30-day mortality may have been small.

B. Model Performance

The logistic regression model showed some promise in predicting 30-day mortality. Although the logistic regression's accuracy (0.64) was worse than the accuracy we would have had if we had predicted that everyone would live (0.88) our model did have fairly high sensitivity and specificity, which suggests that it may perform well in detecting which patients might need advanced care and giving doctors opportunities to do early interventions.

VI. Limitations

Like with any investigation, the results obtained in this study should be considered in light of various limitations. One such restriction comes from the incomplete knowledge of event times due to loss to follow up for certain patients. In our Cox survival model, we were forced to assume that individuals who were lost to follow-up had the same hazard of death as those who remained in the study. In our logistic regression model, this led us to only consider 30 days of follow up. As a result, the predictions of our logistic regression model cannot be applied to follow up periods that exceed 30 days.

In addition, our sample size for our analyses were relatively small (n = 294 for our logistic model and n = 299 for our Cox model). Although the acknowledgement of Small Sample Size (SSS) has become a cliché in research, it is still worth mentioning given the impact of this single number on statistical power. The lower *n,* in addition to the fact that patients were only taken from two hospital sites, also limits the generalizability of our findings beyond the population involved in the study. A larger, more diverse sample would have allowed us to predict those who may die after heart attacks in the general population. Potential extensions to address these issues are presented in the next section.

VII. Future Scope

Further research is warranted and could aid in the prediction of death following a heart attack. It is clear that recruiting more people is important; however, future work could also consider gathering data from individuals outside of Pakistan. Culture and ethnicity both influence lifestyles and make individuals more or less likely to develop

heart failure. It therefore may be useful to seek patients from many cultures and countries to simultaneously diversify the sample and the generalizability of the findings.

Another potential avenue would be to include other variables pertaining to physical health and family history. A discrete variable, say ‘exercise’, could help categorize individuals based on how many days a week they engaged in rigorous physical activities. The relative strength of heart muscles and relative endurance may play an important role in patient survival. Future researchers could also survey patients on their ancestry to see if members of their family experienced heart failure or a heart attack, and ask how many months their relatives lived for post-event. Further, on the recommendation of our subject matter expert, future researchers would likely have more success at predicting outcomes with information on drug therapies administered to each patient. Having more covariates would ultimately increase the predictiveness and explanatory power of many popular methods, including LASSO logistic regression and the Cox survival model. In the end, there exist many possibilities for how this work can be extended and offer a more granular view into the underlying factors affecting survival.

VIII. Appendix

Code is available at links below:

A. Dataset
B. Logistic Regression Analysis
C. Survival Analysis

IX. Supplementary Materials

Figure 1

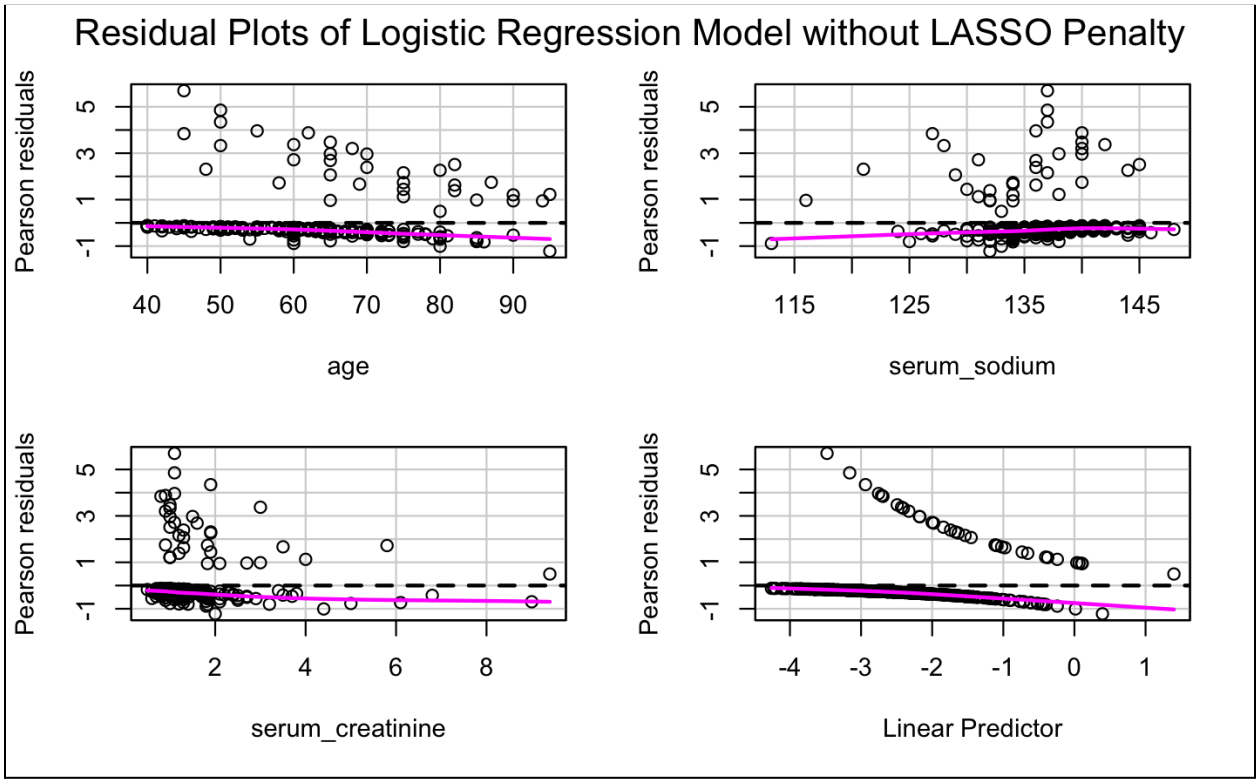


Figure 2

Influence Plot of Logistic Regression Model without LASSO Penalty

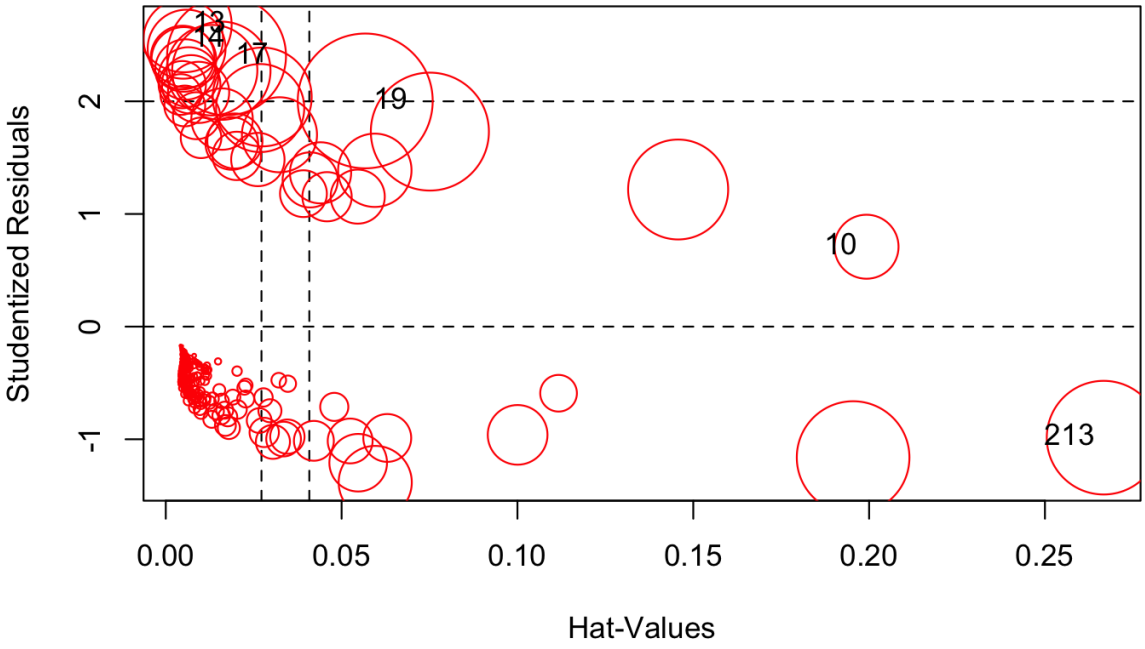


Figure 3

Martingale Residuals of Null Cox Model for Ejection Fraction