Beating the Odds:
A Statistical Analysis to Explore Prognostic Factors for Surviving a Heart Attack

STA 4903: Applied Survival Analysis

Abby Mueller (srb567), Rani Misra (frz630), Carroll Montgomery (cqw575)

Spring 2024

INTRODUCTION

Cardiovascular diseases, especially acute myocardial infarction (MI), continue to present significant global health challenges, contributing substantially to morbidity and mortality rates. According to Brindles Lee Macon at *Healthline*, acute myocardial infarction, commonly known as a heart attack, occurs when blood flow to the heart is abruptly cut off and causes tissue damage. The cut off is typically caused by a blockage of the coronary arteries. This occurs after a buildup of plaque that contains fat, cholesterol, and cellular waste, potentially inducing a blood clot can further restrict flow. The coronary arteries are a crucial channel of supply of oxygen-rich blood to the heart. When it is blocked, blood flow to the heart is significantly decreased or stopped completely, causing a heart attack.

Common symptoms of MI include sustained pressure in the chest, pain in the chest, back, jaw, and other areas in the upper body, shortness of breath, sweating, vomiting, and anxiety. It is to be noted that women may not present these typical systems when it comes to MI. They are more likely to feel jaw pain, upper back pain, flu-like symptoms or no pain at all. As MI is a more common issue in men, knowledge about female-specific symptoms is less advanced, and many conclusions are based on research conducted on men (Jackson et al., 2000). Our analysis further explores the differences between the two sexes.

People with fatty diets are prone to heart diseases. High levels of triglycerides, low-density lipoprotein, and cholesterol are all indicators of a risk of heart diseases and potentiality of MI. Additional risks include high blood pressure, high blood sugar levels, smoking, age, sex, family history of cardiovascular diseases, stress, physical inactivity, drug use, and preeclampsia. Men are at a higher risk of getting a heart attack after age 45 and women after age 55, with the majority of heart attacks occurring in men.

Treatments for this condition may include percutaneous coronary intervention, or an angioplasty with a stent, in which the clogged artery is cleared to allow for the flow of blood to the heart. A coronary artery bypass graft might also be considered to restore blood flow by rerouting veins and arteries so that blood can move around the blockage. A general bypass could also be used to create new passages for blood to flow to the heart. Additionally, several medications can also be used to treat MI. These include blood thinners to break up blood clots and improve blood flow through narrowed arteries, or antiplatelet drugs to prevent new clots from forming and current clots from growing. Nitroglycerin can also be prescribed to widen blood vessels, beta blockers can lower blood pressure and stress to the heart in general, and diuretics may be used to decrease fluid buildup and the heart's overall workload.

Recovering from MI depends on the amount of damage to the heart. If the damage is substantial, the heart may be unable to pump a sufficient amount of blood throughout the body, which could result in heart failure. Additionally, the risk of valve problems resulting in abnormal heart rhythm or arrhythmia is increased.. It is estimated that one in five people who have had a heart attack have a second heart attack within five years. Therefore, following up with medical professionals or undergoing a cardiac rehabilitation program along with medication and a healthy diet is recommended.

In response to these global health challenges, the Worcester Heart Attack Study, led by the Department of Cardiology at the University of Massachusetts Medical School, serves as a pivotal investigation aimed at uncovering the complex factors that influence post-MI survival rates. Spanning thirteen one-year periods from 1975 to 2001, the Worcester Heart Attack Study (Spencer et al., 1999) conducts a comprehensive and longitudinal examination of MI patients

admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. This study's dataset includes the following key variables:

- id – ID code from 1 to 100,
- admitdate – admission date in mm/dd/yyyy,
- foldate – follow up date in mm/dd/yyyy,
- los – length of hospital stay in days,
- lenfol – follow up time in days,
- fstat – follow up status with 1 = dead and 0 = alive,
- age – in years,
- gender - 1 = female and 0 = male,
- bmi – body mass index in kg/m^2,

providing a novel perspective on the multifaceted nature of post-MI prognosis. The data relies on collection of accurate follow-up status from patients, potentially for years after the initial hospital stay. However, some patients are necessarily lost for a variety of reasons, whether they move away, choose to leave the study, or simply can no longer be reached. These patients still provide some information on survival time before being lost, so their data was kept, but we consider their survival as censored, as we do not know their exact time of death.

As previously mentioned, cardiovascular disease is the leading cause of death in the United States. Not only is CVD a growing issue as the nation ages, but the resulting costs of CVD are raising insurance premiums for all Americans. 15% of all health care costs come from managing and treating CVD (Birger et al., 2021), highlighting the importance of understanding CVD from a financial perspective. Furthermore, the average American diet includes a high caloric intake, increasing the chances of developing heart disease and myocardial infarction. While American's dietary trends fail to move away from their unhealthy habits, this analysis furthers the study of which factors play the greatest role in predicting survival times following an MI event.

OBJECTIVES

The primary objective of our study is to assist healthcare professionals in identifying higher-risk patients. Our model seeks to enable doctors to pinpoint patients with a higher risk profile, potentially leading to improved survival outcomes. Leveraging advanced statistical methodologies, our research seeks to offer actionable insights for both medical professionals and patients. This deeper understanding of the trends and critical factors influencing survival outcomes across diverse patient characteristics can aid in tailoring personalized care pathways.

Although doctors and heart attack specialists are experienced in the biological issues that cause heart attacks, we found that most professionals are either misled by health statistics or outright unaware of the benefits a skilled analysis can provide. This issue is common to "doctors, patients, journalists, and politicians alike" (Gigerenzer et al., 2007), creating a "collective statistical illiteracy". In other words, an uneducated culture is in place surrounding statistics in the medical landscape. Our analysis intends to demonstrate the profound utility of survival regression models. We quantify medical data to predict survival times and identify trends that may escape the notice of industry professionals. Through the combined efforts of statisticians and doctors alike, we hope that survival data can ultimately be understood to benefit patients and improve patient treatment methods in the future.

Given the intricate interplay between demographic, clinical, and lifestyle factors in post-MI prognosis, our study takes a multifaceted approach. We delve into the complex relationships between age, gender, BMI, and survival rates, recognizing the risks within patient

populations. This approach advocates for personalized care strategies tailored to individual risk profiles. Doctors who understand the exact figures and relationships among the data can combine their knowledge with their medical background to form further conclusions and improve treatment methods for each patient.

MODELS AND METHODS

Before conducting any statistical analysis, extensive preprocessing steps were undertaken to ensure data quality and consistency. We explored variable transformation methods, such as a logarithm, standardization, and polynomial transformation, but found that no additional changes were necessary. We did convert the given survival time variable, lenfol, from a unit of days into months, but this transformation is arbitrary. We proceeded to form two models: the Cox Proportional Hazards (PH) model and Accelerated Failure Time (AFT) model to explore survival times post-MI.

Cox's Proportional Hazards (PH) Model with Stepwise Regression:

Initially, we constructed a full survival regression model incorporating all predictor variables from the dataset. Subsequently, a stepwise regression technique was employed to identify significant predictors influencing the response variable of follow-up months post-MI. This iterative process of variable selection yielded a reduced model, elucidating the primary covariates impacting survival outcomes. Stepwise regression involves first including all variables in the model, then individually removing variables that are not statistically significant. After each elimination, the model is refit and assessed again under common criteria. Variables can also be reintroduced to ensure that no significant variables are missed under a reduced model. The final model required a minimum of 95% significance for each covariate, which resulted in,

$$\widehat{S}(t) = [S_0(t)]^{e^{(0.039 * Age - 0.7 * BMI)}}$$

Here, the baseline survival function was assumed from Kalbfleisch-Prentice or Breslow estimates, with the Kalbfleisch-Prentice estimates of the survival function used as a baseline of the average of all predictors, including 95% confidence limits over time.

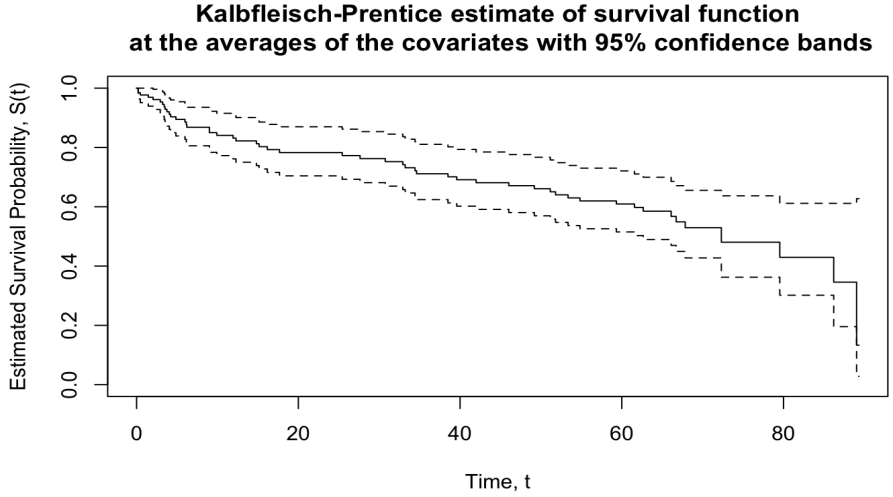

Figure 1: Kalbfleisch-Prentice baseline of the survival function for the Cox Proportional Hazards model

Accelerated Failure Time (AFT) Model:

Simultaneously, AFT models were constructed using various lifetime distributions to assess their efficacy in predicting survival times post-MI. Age, gender, BMI, and an age-gender interaction term were incorporated into the AFT models to capture the complex relationships between these variables and survival rates. Unlike the Cox Proportional Hazards model, AFT models require selection of an underlying distribution. Distributional selection was carried out through comparison of log-likelihood of survival time of the Exponential, Weibull, Lognormal and Loglogistic distributions. After observing very similar log-likelihoods, we compared the significance of the additional parameters estimated in more complex distributions. We found the parameters to be insignificant, and lacking a considerable difference in log-likelihood, we deferred to a more parsimonious model involving the fewest parameters, the exponential distribution.

Distribution	Parameters	Log-Likelihood
Exponential	1	-284.3
Weibull	2	-281.05
Lognormal	2	-282.86
Generalized gamma	3	-280.03
Loglogistic	2	-282.17

Table 1: Log-likelihoods for each distribution in the Accelerated Failure Time model

As far as variable selection, the model was evaluated with all variables for the Exponential, Weibull, Lognormal, Gamma and Loglogistic distributions. Variables that were significant for the majority of distributions were selected for the final exponential model under a minimum of 95% confidence. Insignificant predictors were refitted and removed as necessary. The AFT model under the exponential distribution yielded more significant parameters than the previous PH model, including an interaction term, as represented by:

$$\widehat{S}(t) = e^{-t * e^{(5.87 - 0.5 * Age - 3.93 * Gender + 0.09 * BMI + 0.05 * (Age * Gender))}}$$

	Value	Standard Error	Z	p
Intercept	5.874	1.62	3.63	<.001
Age	-0.053	0.015	-3.39	<.001
Gender	-3.93	1.81	-2.17	0.028
BMI	0.094	0.038	2.49	0.013
Age-Gender interaction	0.05	0.024	2.06	0.039

Table 2: Values of each coefficient, including standard error and significance levels for the Accelerated Failure Time (AFT) model

Model Evaluation and Comparison

After constructing the models, rigorous evaluation and comparison were conducted to compare their performance and predictive accuracy. Specifically, we evaluated individual coefficients, significance levels for each regressor, hazard ratios based on one-unit increases, and Kaplan-Meier estimates of survival functions stratified by gender as pivotal components in the model evaluation process.

Cox Proportional Hazards (PH) Model Diagnostics

The first diagnostic test we performed on the PH model was calculation of the deviance residuals. This test relates the expected event time to what was observed for each observation. These residuals are calculated and a smoothed line with confidence intervals is fit for the mean of the residuals. Ideally, this line should fall at zero, and in our case the line does stay close to zero. Some observations have residuals as high as two, but the important conclusion from this test is that the data does not show a mean that is statistically different from zero. The slight increase in the fitted line near the ends is not substantial enough to cause concern.

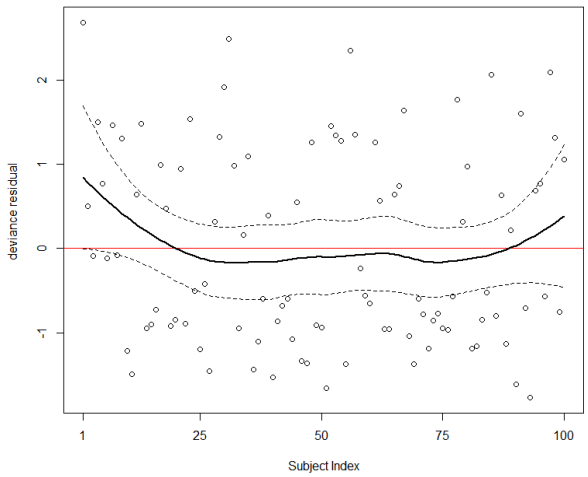

Figure 2: Plot of deviance residuals for each subject observation, with a fitted line around the mean and 95% confidence intervals

The next diagnostic plot we produced for the PH model was the Schoenfeld residuals. They are similar to deviance residuals insofar as they compare expected times to observed, but Schoenfeld residuals exist only for uncensored observations. They also are calculated for each covariate individually, allowing us to isolate a potentially problematic predictor. Like deviance residuals, Schoenfeld residuals test the PH assumption and can also identify outliers with large values compared to the main cluster. Our Schoenfeld residuals suggest that the fit may not be perfect, but the proportional hazards assumption is not violated. As before, an increase in the smoothed fit for large time points is challenged by the lack of local data.

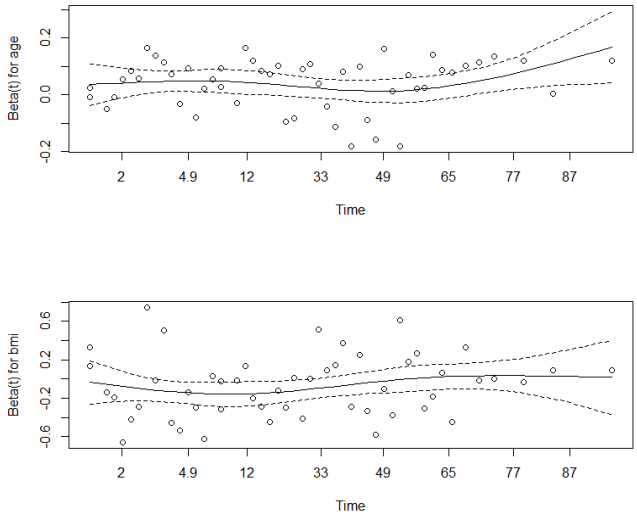

Figure 3: Schoenfeld residuals for uncensored observations in the PH model, stratified by covariate

The final diagnostic test we performed on the Cox Proportional Hazards model was a plot of the Cox-Snell Residuals, which are an estimate of an observation's cumulative hazard. Each residual should have a unit exponential distribution, so their plot versus the estimated cumulative hazard would fall on the identity line. Our Cox-Snell residuals demonstrate a marked downward shift from the identity line, particularly for larger residuals. However, almost half of our data is censored. According to *Statistical Methods for Survival Data Analysis*, extended Cox-Snell residuals are underestimated for censored observations. Even after an adjustment by adding to each censored residual, the fit remains far from ideal. Based on our previous tests, we cannot assume that the PH assumption is violated. Although the overall model fit is not optimal, the Cox-Snell residuals only reflect a partial issue with the Proportional Hazards model.

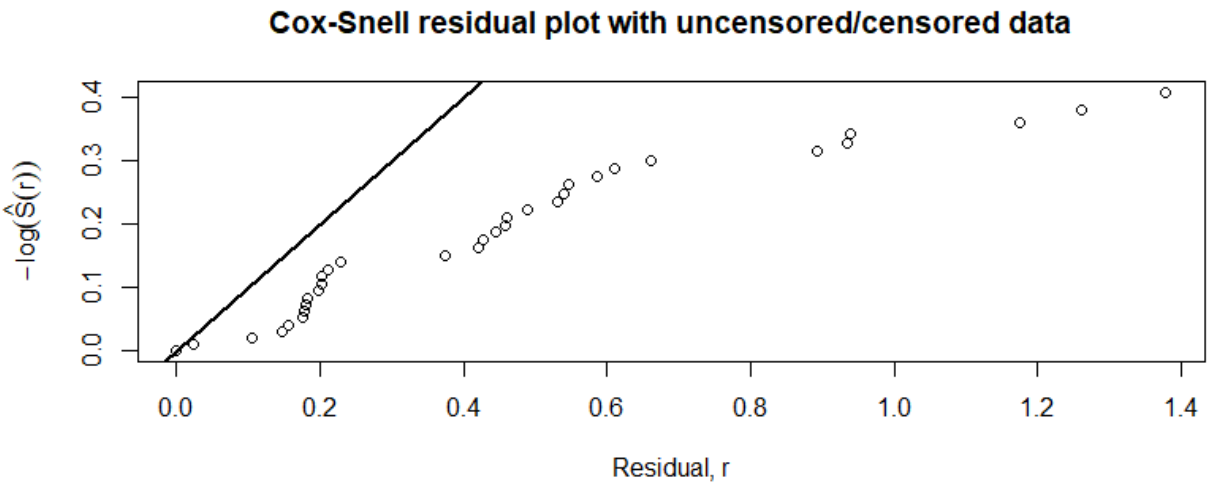

Figure 4: Extended Cox-Snell residuals for the Proportional Hazards Model against estimated Identity Line fit

Accelerated Failure Time (AFT) Model Diagnostics

As a first test of the AFT model, a cumulative hazard plot was generated over time, providing insights into the trajectory of hazard rates post-MI. The linearized cumulative hazard over time should fall along a straight line fit, as the hazard for the exponential distribution is constant. In our case, the line has a steeper slope for times 20 to 60 where the cumulative hazard is increasing at a faster rate. This indicates that our λ may not be entirely constant, but the overall fit ($R^2 = .95$) is acceptable for our use.

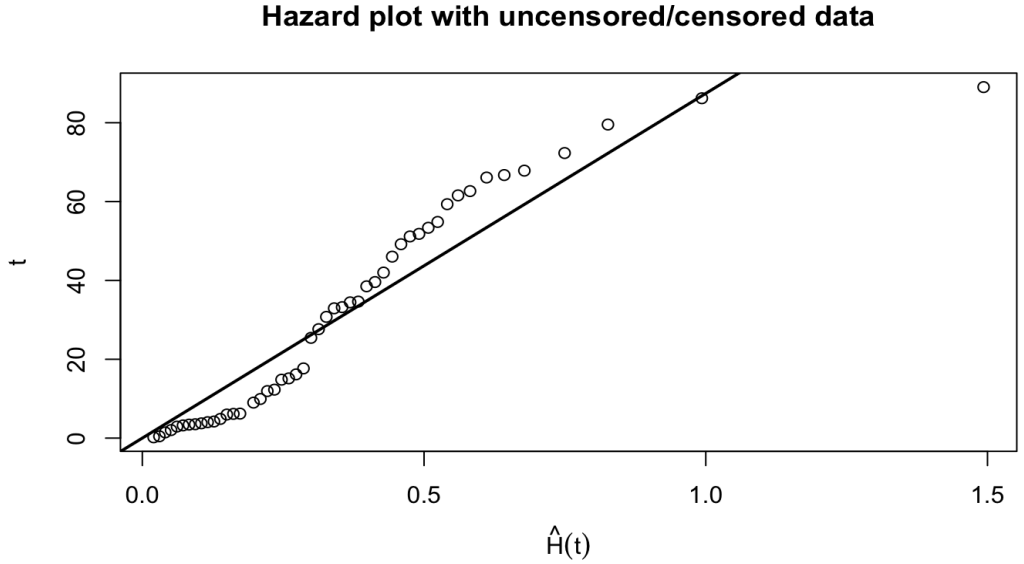

Figure 5: Plot of the linearized hazard for each data point over time, including an overlay of the ideal fit

The other diagnostic test we performed for the Accelerated Failure Time model was a plot of the Cox-Snell residuals. Like the Proportional Hazards model, the fit of the Cox-Snell residuals to the unit exponential line is not ideal. However, the significant percentage of censored observations partially explains the deviation. Since the plots are similar for both models, we cannot compare overall model fit using the Cox-Snell residuals.

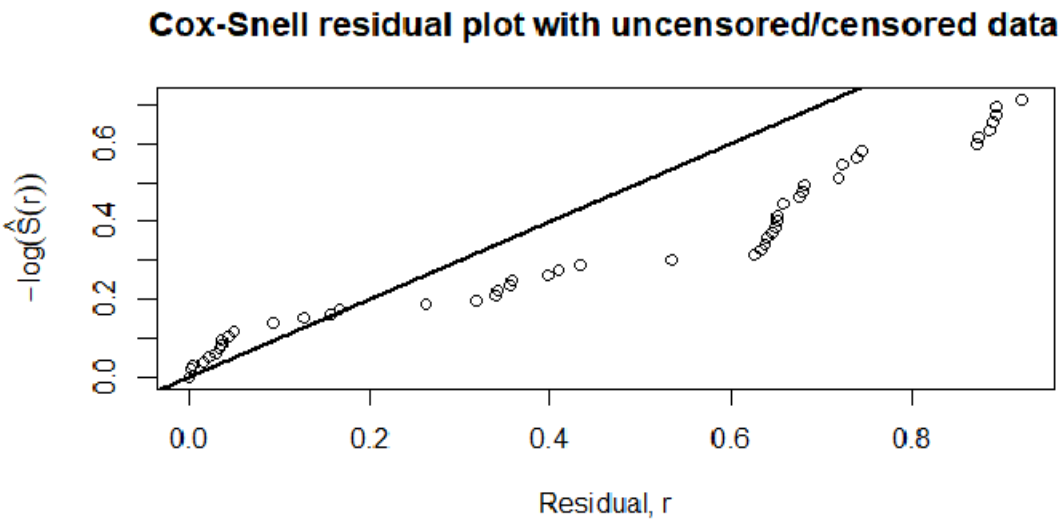

Figure 6: Cox-Snell residual plot for the Accelerated Failure Time (AFT) model

ANALYSIS RESULTS

We have already detailed our model formulation techniques and final equations, but would like to evaluate further on a few critical findings that result from the models. First, our methods utilized two models, both the AFT and Cox PH. We would like to compare them and identify which model provides a better fit to the data. We calculated the log-likelihoods, an indicator of the joint probability of each observation falling under the model, as well as AIC and BIC. Simple log-likelihood will increase even after including insignificant predictors, while AIC and BIC add penalties for additional covariates. We found the Cox Proportional Hazards model to have an average increase of 75 between the three comparison methods. Therefore, the PH model is more accurate and provides a better fit to the data, even though it uses fewer predictors.

Model	Log-Likelihood	AIC	BIC
Exponential AFT	-270.17	-280.17	-281.68
Cox PH	-197.05	-201.05	-201.65

Table 3: Comparison of log-likelihood, AIC and BIC for both models

Additionally, the coefficients of each model require further explanation. In particular, unusual coefficient signs for BMI are of concern. For the PH model, BMI has a negative coefficient, indicating that an increase in weight will increase estimated survival time. In the other model, BMI has a positive coefficient, but both models reflect the same trend. The Proportional Hazards model has the regression equation in the exponent of the baseline survival function. The exponentiation of the BMI coefficient is 0.93, and the baseline survival is always between zero and one, meaning that an increase in BMI will increase the survival probability. In the simpler case of the AFT model, the positive coefficient on BMI directly increases the estimated survival. The same logic can be extended for the Age coefficient, which also has the same effect in both models.

It is reassuring that both models provide similar information from the BMI covariate, but the result is uncommon. In most cases, people with larger weights do not live as long, especially after a heart attack. The small sample size and high censoring in our data provides some explanation for this result. More importantly, our data was likely skewed toward patients who are underweight, in which case those with a healthy weight would live longer. Patients with extreme BMI would not live after suffering a heart attack and would then not be considered in this study. Although our finding is uncommon, it is a simple function of the data, and should not be extended for the entire study population.

Next, we generated hazard ratios for the Cox Proportional Hazards model. This model assumes that the ratio of hazards for a one-unit increase in a covariate will be constant over time when all other covariates are kept constant. For example, in our case, a one-unit increase in BMI will cause an estimated 7% decrease in hazard (log scale), with a 95% confidence interval between 0.87 and 1.00. Because the confidence interval includes 1, we cannot assume that BMI is statistically significant at the 95% level. A one-unit increase in Age increases the estimated hazard by 4%. In this plot, the hazard ratios are exponentiated from the original model form.

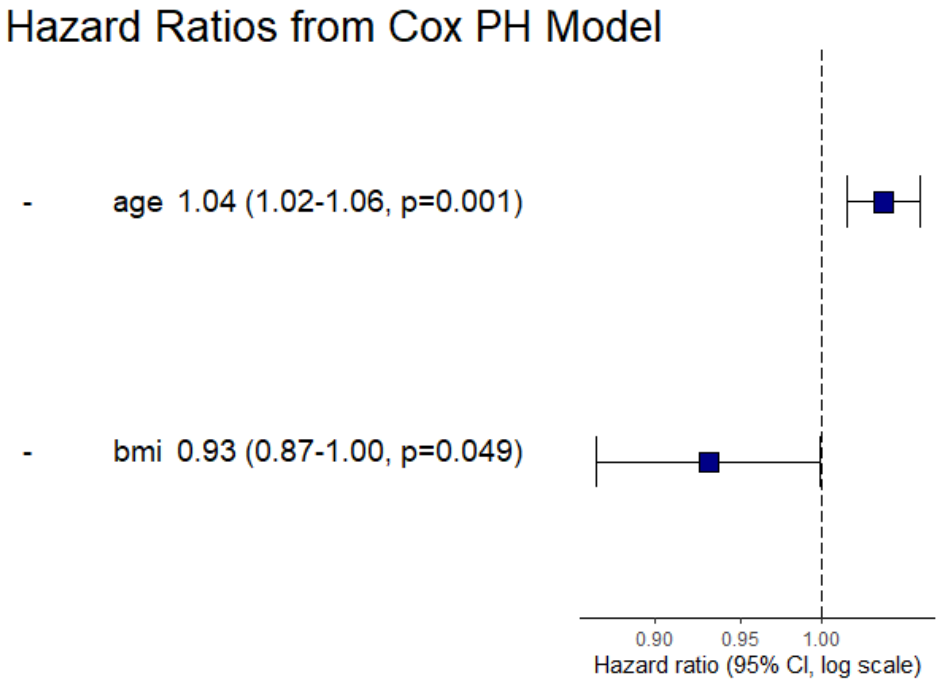

Figure 7: Hazard ratios for a one-unit covariate increase in the Proportional Hazards model

Next, we stratified the data by gender and compared the graphs of the Kaplan-Meier nonparametric survival estimates. Females have a lower survival probability for almost all time points. The use of an age-gender interaction term in the AFT model provides further insight into this difference. The coefficient on this term directly cancels the original age term, so females have a lower survival probability in general, but no significant correlation with age. This can be corroborated by the shortage of studies geared towards women that could potentially cause a lack of medical attention towards their cardiovascular health in general, leading to a lower survival rate.

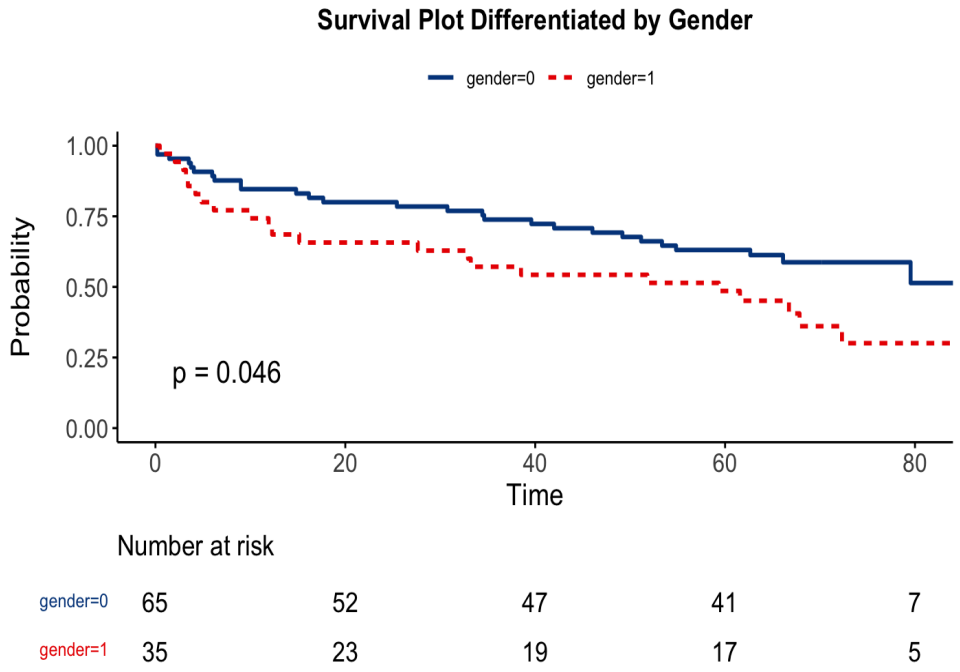

Figure 8: Plot of the Kaplan-Meier nonparametric survival estimates for each gender and numbers at risk over time

CONCLUSION

Our analysis has illuminated crucial insights into the prognostic factors influencing survival rates among patients following acute myocardial infarction (MI). Through the application of advanced statistical techniques such as Cox's proportional hazards (PH) model and accelerated failure time (AFT) model, we have isolated nuanced relationships in post-MI survival.

The identification of age and body mass index (BMI) as primary risk factors that were included in both models underscores the necessity for individualized prognostic assessments. The Cox PH model's emphasis on these factors alone sheds light on the unique interplay between patient demographics and post-MI outcomes. Our inclusion of an interaction term in the AFT model further stratifies the treatment methods between women and men as a response to MI.

We have detailed the strengths of our models as well as their shortcomings, but we should also discuss their intended scope. Low sample size and high censoring limit the ability for these models to be adequately scaled to a larger data set. The predictive ability of this model may be accurate for the fitted data, but outside samples should be compared and evaluated independently. Covariates that deviate from general averages, such as BMI, also misrepresented their true relationship in long-term survival time. These models fit this data, but may not work well for modern data collected from other sources.

Our findings not only contribute significantly to the scientific understanding of MI prognosis but also hold profound implications for clinical practice. By recognizing age, BMI, and gender as pivotal determinants, healthcare providers can tailor interventions and treatments more precisely, potentially leading to enhanced patient outcomes and improved quality of life.

However, our study also underscores the ongoing necessity for further research and exploration. Investigating the combined effects of specific treatments, medications, or lifestyle

interventions with these prognostic factors could offer deeper insights into optimizing post-MI care strategies. Additionally, longitudinal studies tracking patient outcomes over extended periods would provide invaluable data on the long-term impacts of interventions on survival rates and quality of life. In the end, our study is carried out from a purely statistical perspective. Our expertise is in analysis, not medical diagnosis, and we lack the industry background of experienced medical professionals. We hope to bridge the gap between statisticians and doctors, in order for both groups to understand the qualitative information behind medical data and make critical insights to improve treatments across the board.

In essence, our research underscores the pivotal role of rigorous statistical analysis in uncovering critical prognostic factors, thereby paving the way for targeted interventions and personalized care pathways in managing acute myocardial infarction. This holistic approach not only benefits individual patients but also contributes to the broader advancement of cardiovascular healthcare practices.

REFERENCES

Birger, M. & Kaldjian, A.S. & Roth, G.A. & Moran, A.E. & Dieleman, J.L. & Bellows, B.K. (2021, July 27). Spending on Cardiovascular Disease and Cardiovascular Risk Factors in the United States: 1996 to 2016. *American Heart Association 144*(4), 271-282. https://doi.org/10.1161/CIRCULATIONAHA.120.053216

Gigerenzer, G. & Gaissmaier, W. & Kurz-Milcke, E. & Schwartz, L. M. & Woloshin, S. (2007, November). Helping Doctors And Patients Make Sense of Health Statistics. *Association for Psychological Science (8)*2, 53-96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Jackson, D. & Daly, J. & Davidson, P. & Elliot, D. & Cameron-Traub, E. & Wade, V. & Chin, C. & Salamonson,Y. (2000, June 29). Women recovering from first-time myocardial infarction (MI): a feminist qualitative study. *Journal of Advanced Nursing 32*(6), 1403-1411. https://doi.org/10.1046/j.1365-2648.2000.01622.x

Lee, E. T. & Wang, J.W. (2013). *Statistical Methods for Survival Data Analysis* John Wiley & Sons.

Macon, Brindles Lee. "Myocardial Infarction (Heart Attack): Symptoms and More." *Healthline*, Healthline Media, 3 Nov. 2021, www.healthline.com/health/acute-myocardial-infarction.

Spencer, F. A. & Meyer, T. E. & Goldberg, R.J. & Yarzebski, J. & Matton, M. & Lessard, D. & Gore, J.M. (1999). Twenty Year Trends (1975-1995) in the Incidence, In-hospital and Long-Term Death Rates Associated With Heart Failure Complicating Acute Myocardial Infarction. *Journal of the American College of Cardiology 34(5)*, 1378-87. https://doi.org/10.1016/S0735-1097(99)00390-3