COMP 152: Statistical Bioinformatics in R

Project Paper:

Survival Analysis of Heart Attack Patients using Echocardiogram Data

Sophia Zhang

December 3, 2020

Introduction

Myocardial infarctions, more commonly known as heart attacks, is a serious health concern with 3 million cases per year in the US alone (Benjamin et al., 2019). Heart attacks are a leading cause of death among adults worldwide and contributes to economic losses and strain on the health care system. A heart attack was ranked in the top five most expensive in-hospital conditions to treat in 2011, with an average cost of $19,000 per hospital stay in the US (Torio, 2013). In many instances, heart attacks are preventable, especially with careful monitoring. Methods to predict the survival rate of the patient is key in allocating time and resources to those who need it most.

The echocardiogram is an ultrasound of the heart that provides moving images from which information on the structure and function of the heart can be determined. A score can be assigned based on the left ventricular segmental wall motion seen in the echocardiogram (a measure of how the segments of the left ventricle are moving). This wall motion score (WMS) can be used as a way to assess the condition of the left ventricle in patients who have or have had a heart attack. The score is the sum of scores of 13 segments of the left ventricle, each assigned a number based on the degree of systolic wall motion abnormality.

In the paper "Short and long term predictive value of wall motion score in acute myocardial infarction", Kan et al. (1986) seeks to predict from certain variables whether or not a patient will survive at least one year after a heart attack. Specifically, the paper uses wall motion score as a predictive value for a patient's mortality. The result of the paper was a single threshold value to predict 1 year survival, which led to a simple and effective prediction tool for this purpose.

The original data analyzed by Kan et al. (1986) are for 345 heart attack patients who had an echocardiogram taken upon admission to the hospital. The available dataset for this paper is a truncated version of the original, with only 132 patient observations. In the data, wall motion score and also wall motion score index is provided for each patient. The wall motion score index (WMSI) is calculated by dividing the score by the number of segments considered in each score (since not all 13 segments is always considered). Wall motion score index is often used instead since it is a better standardized variable than wall motion score. Klein et al. (2009) in their study of 101 patients found the optimal threshold value for predicting mortality was a WMSI of 2.19.

This paper is a combination of reanalysis of the work by Kan et al. and additional original analysis on the truncated dataset. The reanalysis portion involved applying the WMS threshold of 10 (optimal level as determined by Kan et al.) on the truncated dataset and comparing the predictive values. The optimal WMS threshold specific to the truncated dataset was also calculated. New analysis involved predicting probability of survival over time using the Kaplan-Meier estimator method. In addition, new machine learning algorithms such as logistic regression, random forest, and K Nearest Neighbors was applied for binary classification; to predict survival (yes or no) within a year of the heart attack. Age and other clinical data variables was added in the machine learning models. The accuracy of the models in predicting whether patients will die within a year was evaluated.

Project Methods and Data

The data analyzed by Kan et al. (1986) is a sample of 345 unique patient observations taken from the University of Amsterdam hospital during a period of 2 years (1981-1983). Only patients diagnosed with myocardial infarction (heart attack) were considered. This serves to approximate a population with a known history of heart attack. There are 9 explanatory variables in the dataset such as age, wall motion score, and some clinical data. There are 3 outcome variables: alive or not, survival period, and alive at 1 year after hospital discharge. Standardized machines and diagnostic processes were used in data collection. The key variables and their description is displayed in Table 1.

The data available for use is a truncated version of the full data, with 132 patient observations instead of 345. To determine if the truncated data is representative of the original sample, the wall motion score distributions of the survivor and non-survivor groups are compared. The distributions are also plotted using a quantile-quantile plot to assess if they follow an approximately normal distribution. The dataset includes observations for some patients who are still alive but have a recorded survival period of less than 12 months. These were patients who were followed for less than a year. Since the original paper only looks at patients who died within a year or were followed up with in 1 year or more, these censored were removed for the comparison portion of the analysis. Missing data which are not critical (variables other than wall motion score) were replaced with the mean of that variable to preserve as many observations as possible.

Due to the small sample size of the available data, bootstrap resampling is used to determine the 95% confidence interval of the mean wall motion score. Bootstrap resampling is a method where empirical observations are randomly sampled with replacement to assess the parameters of a population. It is applicable here since the available dataset is only a fraction of the true dataset or population, and as such there is some uncertainty in the mean. The sample is judged to be not representative of the full dataset if the 95% confidence interval does not encompass the mean stated in the paper.

Kan et al. hypothesized that a wall motion score can be an independent prognostic of mortality based on knowledge of its correlation with infarct size (directly correlated with mortality). In other words, there should be a distinct difference in WMS between survivors and non survivors. The researchers used the Student's t-test to quantify differences between the groups. Therefore, this paper also hypothesizes that there is a statistically significant difference in the means of WMS in the survivor and non-survivor populations. This paper expects to reject the null hypothesis of equal means in the populations. A formal statistical test in the form of a Student's t-test will evaluate this expectation.

In the original study, the authors determined a wall motion score of 10 is the best threshold for prediction. Based on this model, a patient with WMS equal to or above 10 is predicted to die within a year and a patient with WMS below 10 is predicted to live. This threshold yields the maximum sensitivity and specificity values. These values are based on the proportion of correct predictions (predicted and actual deaths) and is a measure of the efficacy of the predictions. This paper repeats the same calculations for sensitivity, specificity, positive and negative predictive values for the truncated dataset.

A problem the paper raised but did not resolve was predicting the survival time. Knowing how many months or years a patient has to live is very critical for timely treatment. The model in the paper can predict death with some accuracy up to one year on the single WMS variable. As an extension to the analysis this paper generates Kaplan-Meier estimators for the sample as a whole and for groups separated based on a WMS threshold. The Kaplan-Meier estimator predicts the probability of survival over time. This method allows the censored data which is left out of the direct comparison with the original paper to be incorporated, thus maximizing the use of the information in the dataset.

The new analysis in this paper also includes looking at other variables in the data and their correlations, as well as their value in predicting if a patient will live more than 1 year after the heart attack. Supervised machine learning algorithms are natural models to apply since the dataset already includes labels for the patients; labels in this case come from the variable indicating if the patient survived 1 year. Logistic regression, K Nearest Neighbors (KNN), and random forest algorithms are used with the entire dataset, not just the WMS variable, to make predictions for individual patients. The accuracy, sensitivity and specificity values are evaluated for each model.

There is an imbalance in the classification categories; the truncated data contains 92% survivors and 8% non-survivors in 96 patients. Random resampling per the bootstrap method is applied to balance the samples. A combination of oversampling (taking repeat observations for the non-survivors) and under sampling (reducing observations for the survivors group) is used to create a final 50% survivors - 50% non-survivors sample. This is the sample used to train and test the algorithms. A train-test ratio of 70%-30% is used to separate the data into a training dataset for the algorithm to 'learn' from, and a testing dataset to assess the performance of the model and its predictions.

The language and program for this analysis is R, specific packages used include: 'survival' for Kaplan-Meier analysis, 'caret' for machine learning models, 'pROC' for model evaluation metrics, and 'ggplot2' for graphic visualizations.

Table 1. Variable Descriptions

	Variable Name	Description
1	survival	the number of months patient survived (has survived, if patient is still alive). Because all the patients had their heart attacks at different times, it is possible that some patients have survived less than one year but they are still alive. Check the second variable to confirm this. Such patients cannot be used for the prediction task mentioned above.
2	alive	Binary variable that indicates if patient is still alive at the time of follow-up: 0=dead at end of survival period, 1 means still alive
3	age	Age in years when heart attack occurred
4	pericardial effusion	Binary variables, indicates if a Pericardial effusion condition exists, or fluid around the heart. 0=no fluid, 1=fluid
5	fractional shortening	A measure of Contractility around the heart lower numbers are increasingly abnormal
6	E point septal separation (epss)	point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
7	left ventricular end-diastolic dimension (lvdd)	This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
8	wall motion score (WMS)	A measure of how the segments of the left ventricle are moving
9	wall motion index (WMI)	Standardized form of wall motion score. Equals wall-motion-score divided by number of segments seen. Usually, 12-13 segments are seen in an echocardiogram.
10	aliveat1	Derived from 'survival' and 'alive' attributes. Categorical variable to indicate if patient was either dead after 1 year, had been followed for less than 1 year, or patient was alive at 1 year
11	died_1yr	Binary variable for patients followed for more than 1 year, to indicate if patient was either dead after 1 year, or patient was alive at 1 year

Results & Discussion

Data Overview

In this section a brief overview of the data is discussed. The data available is a truncated version of the original, with 132 patient observations instead of 345. The wall motion score distributions of all patients (including those followed for less than 1 year) is presented in Figure 4 below. The mean wall motion score (WMS) of all patients is 14.4 (marked by the black dashed line) with a standard deviation of 5.0. The quantile-quantile plot in Figure 2 shows the actual quantiles and the theoretical quantiles of the data. If the data follows a standard normal distribution, the points should lie perfectly on the black theoretical line. The distribution is approximately normal but skewed right if high WMS outliers are accounted for, indicating that majority of patients have average WMS, but a few rare cases have extremely high WMS.

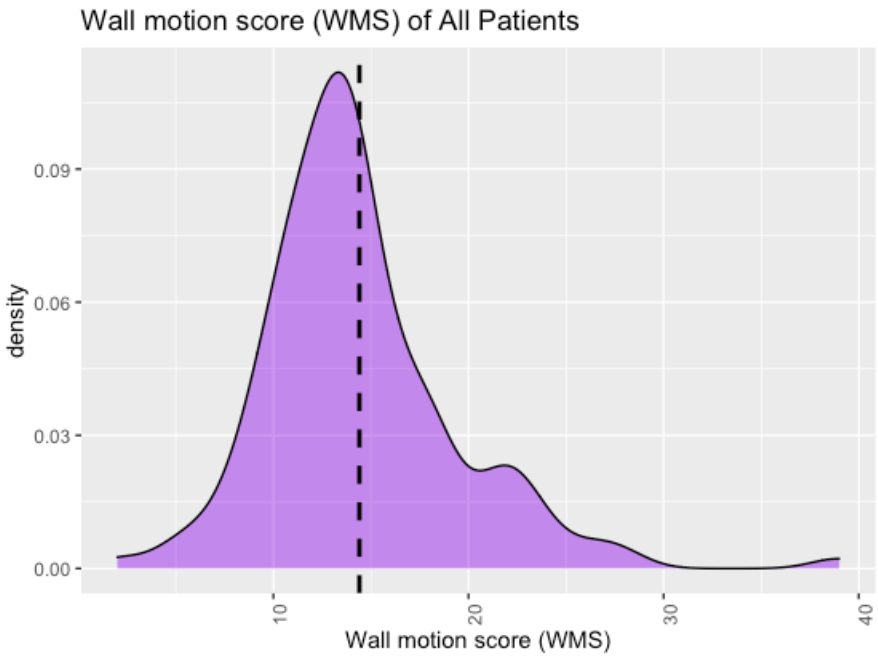

Figure 1. Wall motion score distribution of all patients

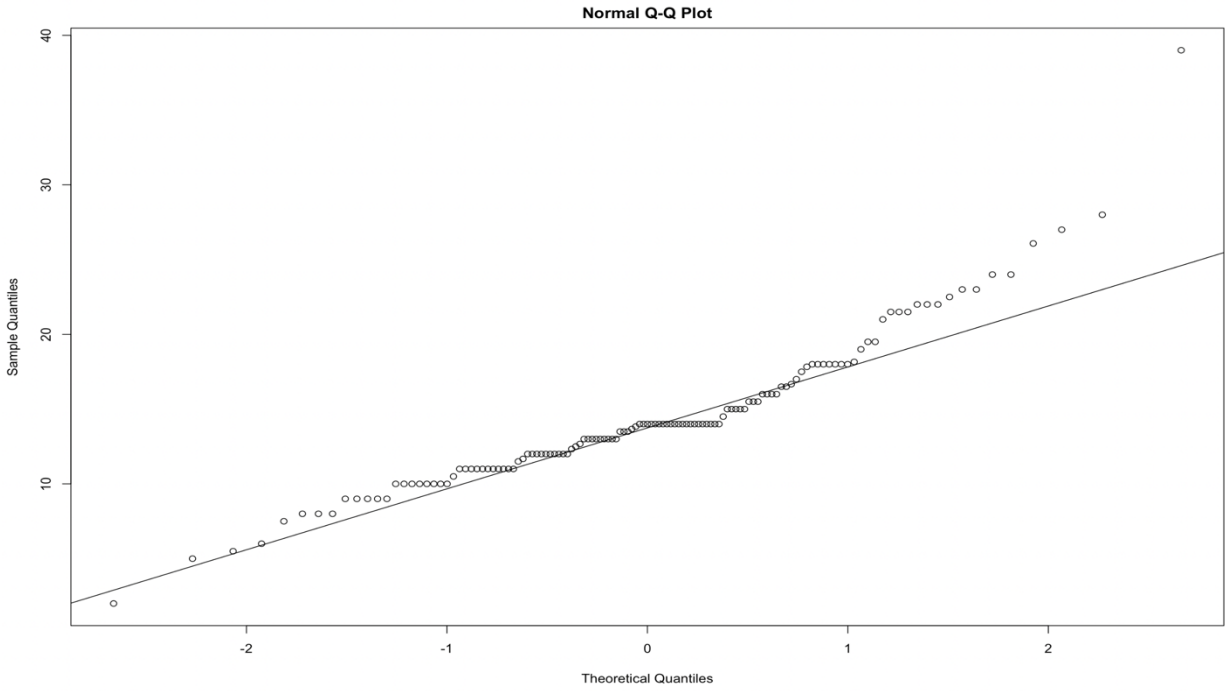

Figure 2. Quantile-quantile plot of Wall motion score of all patients

In the initial data exploration, a mislabeling problem was found. The target variable for whether the patient is alive after 1 year from hospital discharge did not follow the description in the attribute key provided with the data. The target was recalculated based on the survival period variable and corrected. If survival period was larger than or equal to 12 months, then the aliveat1 was marked as 1 (still alive after a year). If survival was less than 12 months and the patient was not alive at the end of that period, then aliveat1 was marked as 0). Some patients were still alive but had a recorded survival period of less than 12 months. These were patients who were followed for less than a year. Since the paper only deals with patients who died within a year or were followed up with in 1 year or more, these were removed for the comparative analysis. Missing data which not critical (variables other than wall motion score) were replaced with the mean of that variable to preserve as many observations as possible.

Comparative Analysis

In this section, a comparison between the original study and this paper is presented. The data for this paper after removing censored data contains 96 patient observations, with 8 non survivors and 88 survivors. Survivor is defined as a patient who is alive more than 1 year after a heart attack. To see if the truncated data is representative of the original sample, the wall motion score distributions of the survivor and non-survivor groups are compared with the original study in Figure 4. A density curve for the patient WMS in the original study is simulated using the mean and standard deviation from the

paper (and assuming a normal distribution) since the original data is not available. The black vertical dashed line is the wall motion score of 10. For the original dataset, the wall motion score of 10 can be used as a threshold to categorize the patients. However, for the truncated dataset, the distinction between the two groups is not so clear and there is a large area of overlap.

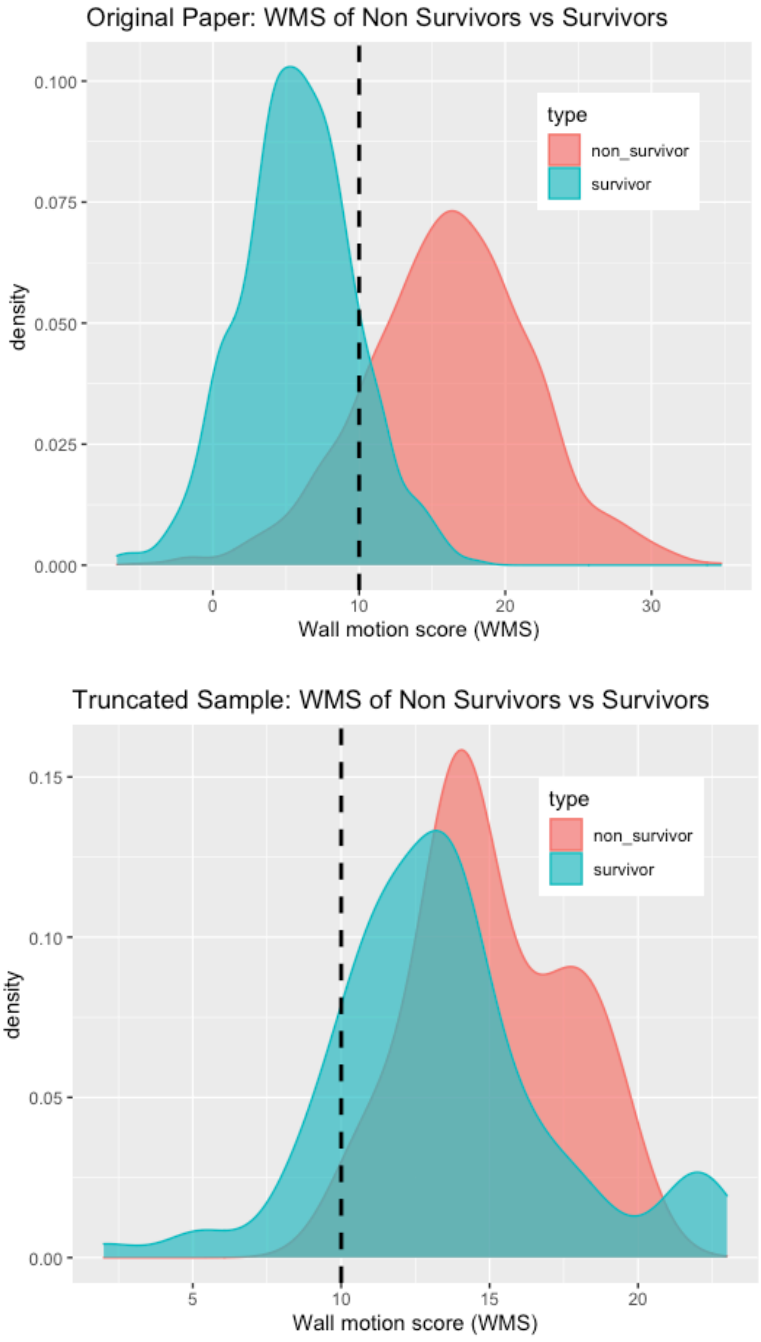

Figure 3. Comparative plots of wall motion score of survivors and non-survivors

Figure 4 shows a box plot for the two groups in the truncated dataset. The data shows that the mean WMS of survivors is less than that of non survivors, the survivors group also has a wider range. The mean WMS of non survivors is 15.1 with a standard deviation of 2.6, mean WMS of survivors is 13.4 with a standard deviation of 3.8. Using bootstrap resampling methods, the 95% confidence interval (CI) of the survivor group mean wall motion score is calculated to be [12.60, 14.17]. The non-survivor group mean wall motion score 95% CI is [13.50, 16.58]. The means of these two groups from the original paper are 5.7 for survivors and 16.2 for non survivors. Assuming the paper means are the population means, only the 95% CI for non-survivors hold the population mean. Therefore, sample is not representative of the dataset in the original paper; an exact replication of the results of the paper cannot be successfully conducted.

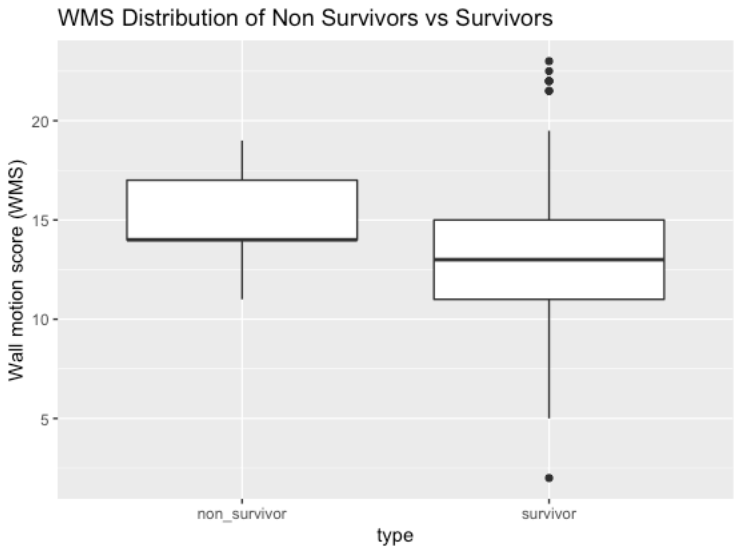

Figure 4. Wall motion score distribution of non survivors group and survivors group

Because sample data is not representative of the original a direct comparison using the same WMS threshold to predict patient survival does not yield the same results as the paper. Figure 5 shows the confusion matrix for non survivors and survivors prediction using a WMS threshold of 10. The predicted numbers of patients that survive or not is displayed on the x-axis, while the actual outcome count is on the y-axis. The original study applied a WMS of 10 to predict that a patient will die within a year (a patient with WMS below 10 is predicted to live).

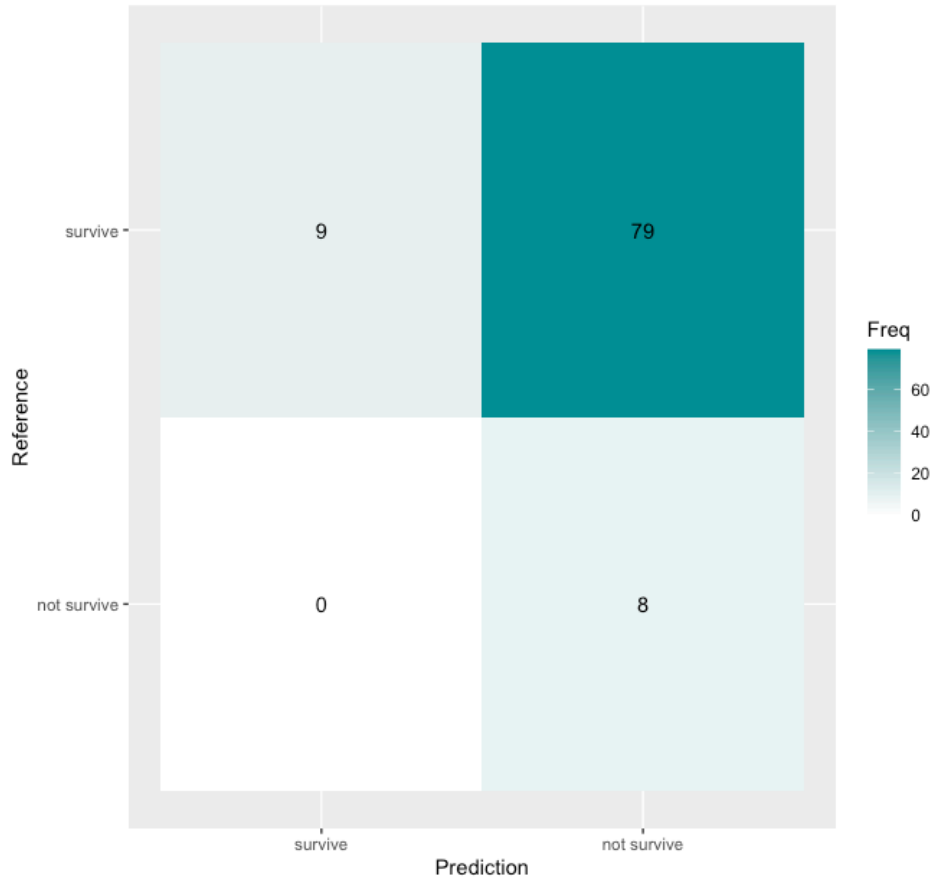

Figure 5. Confusion matrix for non survivors and survivors prediction using WMS threshold of 10

The precision, negative predictive value, sensitivity, specificity, and accuracy of both analyses are compared below in Table 2. The optimum sensitivity and specificity are not achieved at WMS of 10 using the available data. In the paper, precision was low while negative predictive value was high, in this analysis the trend is reversed. Sensitivity, negative predictive value and total predictive accuracy are very poor. Every patient predicted as a survivor will survive, but 90% of the predicted non survivors actually lives. When applied to a different set of data, in this case a subset, wall motion score threshold of 10 may not be a good predictor.

Table 2. Comparison of predictions between original paper and analysis at WMS threshold of 10.

	Paper	Available Data
Positive predictive value (Precision)	61.0%	100.0%
Negative predictive value (TNR)	96.7%	9.2%
Sensitivity (TPR)	88.4%	10.2%
Specificity	85.9%	100.0%
Total predictive accuracy	86.4%	17.7%

Kan et al. hypothesized that there is a distinct difference in WMS between survivors and non survivors, and thus the WMS threshold test can be a valid tool to distinguish between the two groups. In this analysis the Student's t-test is used to formally evaluate this hypothesis on the available data. Based on the distributions of the WMS for the two groups in Figure 4, the WMS of the survivors group is expected to be less than that of the non-survivors group, this is the alternative hypothesis. The null hypothesis is that the two means are equal; there is no difference in WMS between the two groups. Assume independent samples with two unequal and unknown variances based on the original paper. A Welch two sample t-test yields a p-value of 0.061 so at the 5% significance level we fail to reject the null hypothesis of equal means, meaning that we cannot conclude that the survivors group WMS is smaller than the non survivors group WMS.

The p-value for a two sided test (testing if the mean of the survivors group is bigger or larger than the non survivors group) is 0.123 so at the 5% significance level we still fail to reject the null hypothesis of equal means, meaning that we cannot conclude that the survivors group WMS is significantly different than the non survivors group WMS. These results do not support the finding in the paper that there is a significant difference in means between the two groups. This is problematic because the intent of the original analysis is based on primarily using WMS for prediction. Therefore, other variables for prediction needs to be investigated. The intent for the additional analysis in this paper is to explore other predictive methods.

Survival Analysis

This section discusses the predictions for the probability of survival over time using Kaplan-Meier estimators, an extension to the original analysis. The Kaplan-Meier method maximizes the use of the available dataset as it allows censored data to be incorporated. From Figure 6, the probability of a patient living 1 year after a heart attack is 93% and past 2 years is 68%. The probability predicted for

individual groups, using the paper's WMS threshold of 10, is shown in Figure 7. The plot shows time in months on x-axis and probability of survival on the y-axis. The green line is the probability for the high-risk high WMS score group, and the red is for the low-risk low WMS group. There is a large overlap between probabilities if the sample is split this way. Figure 8 shows there is more pronounced difference in survival probability between groups at the WMS threshold level of 13 over the first 2 years. It is interesting to note, the survival probability appears to reverse between the two groups close to year 3, however the error overlaps in this area and it is not conclusive. Figure 9 shows a similar trend when the groups are split by age over or under 60 years old. Younger patients have higher probability of survival in the first two years, but older patients have higher probability of survival if they live past 2 years. Splits using other variables did not yield any conclusive differences. For this limited dataset, an optimal threshold value for prediction can vary and is difficult to determine precisely.

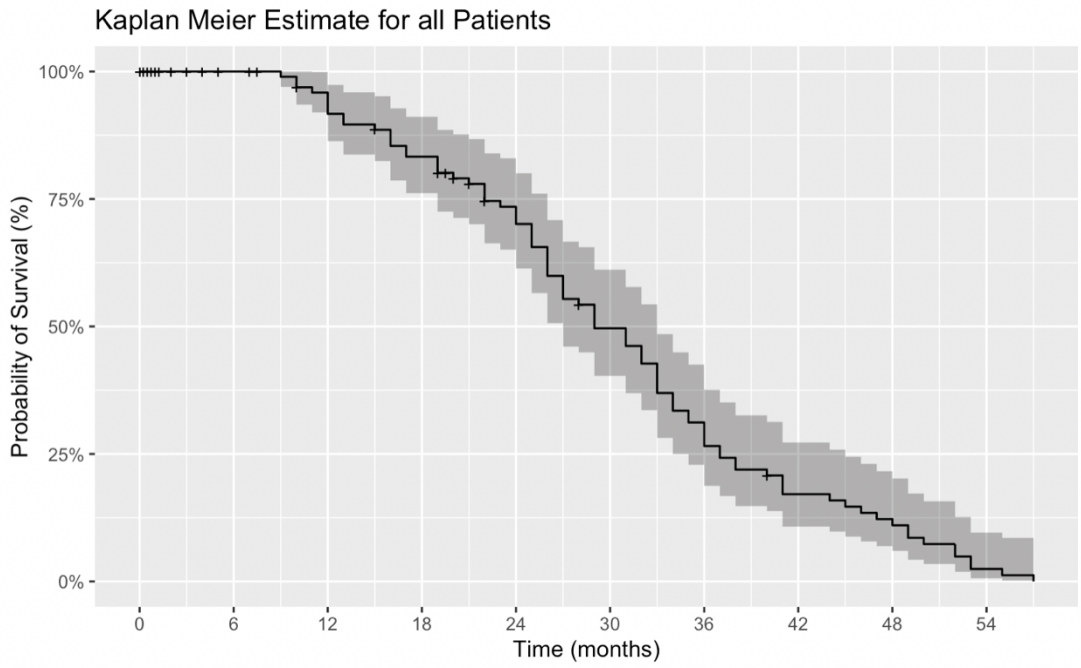

Figure 6. Kaplan-Meier estimate plot for all patients

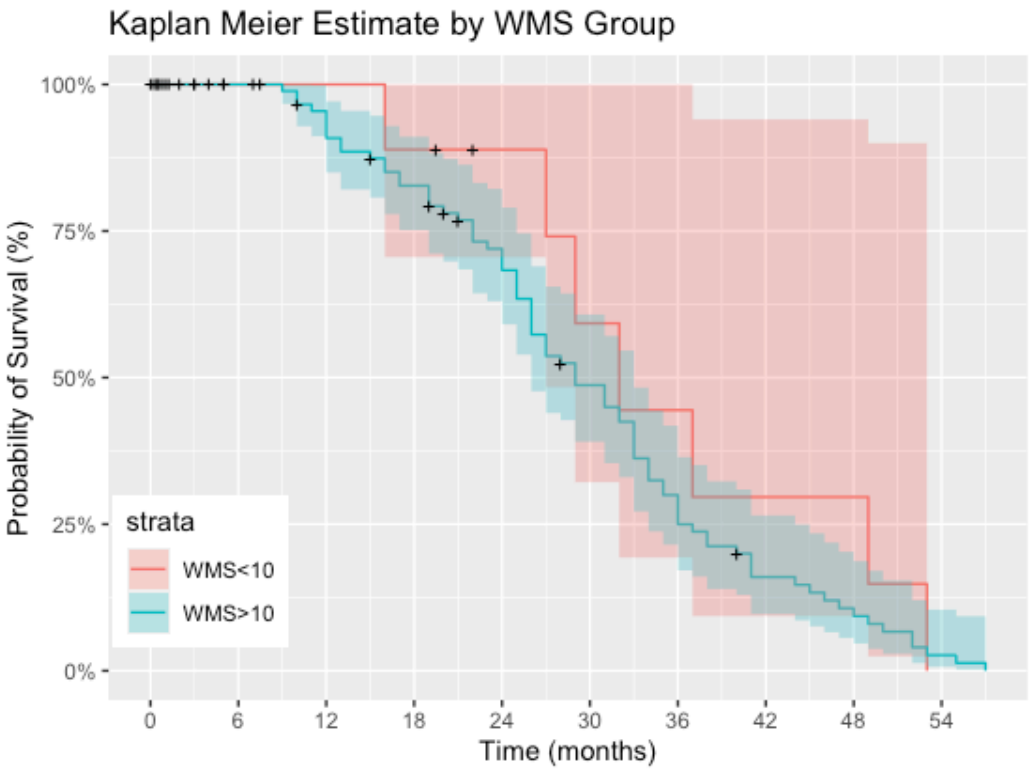

Figure 7. Kaplan-Meier estimate by group for WMS>10 or WMS<10

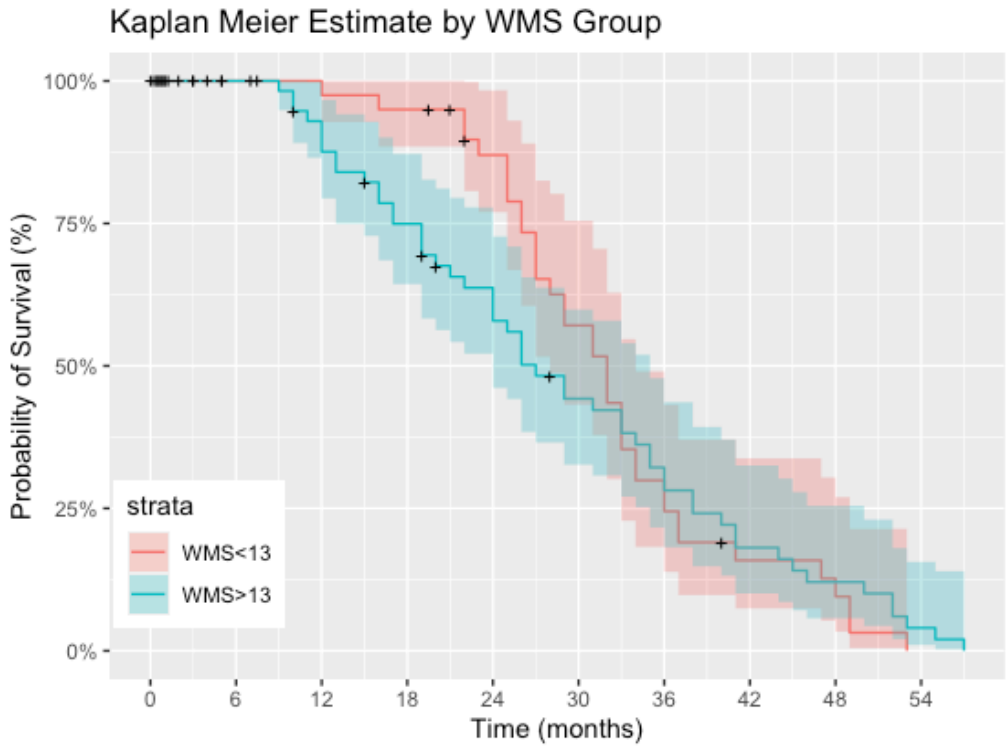

Figure 8. Kaplan-Meier estimate by group for WMS>13 or WMS<13

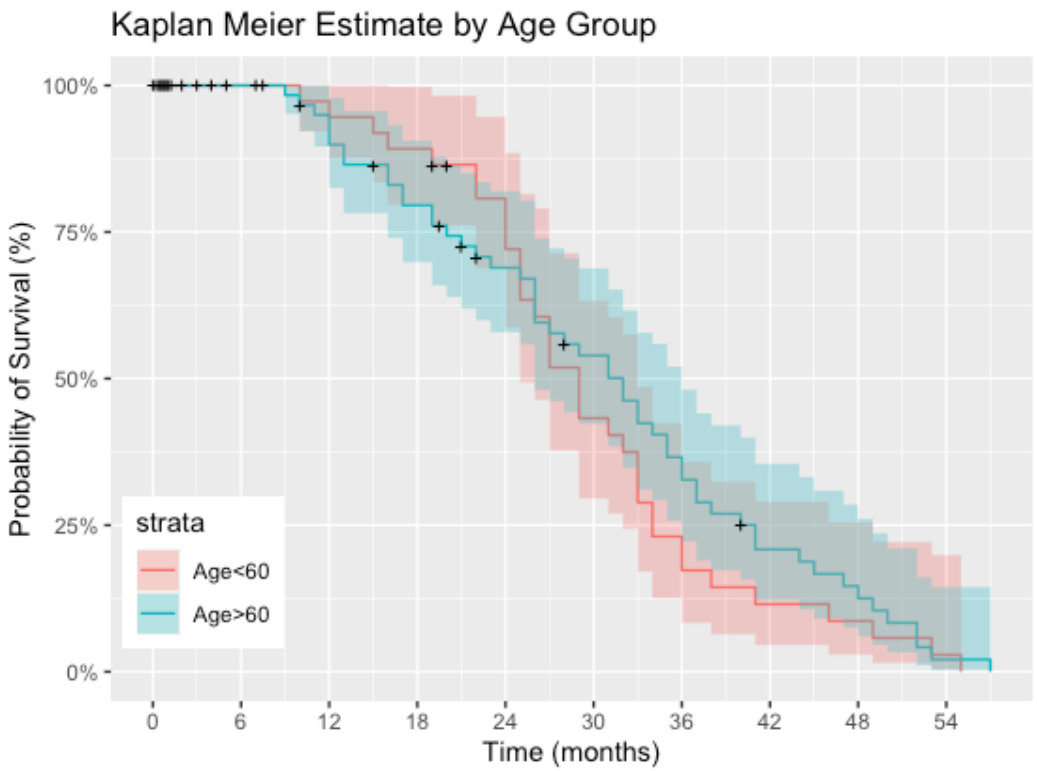

Figure 9. Kaplan-Meier estimate by group for age>60 or age<60

Modeling and Prediction with Machine Learning Algorithms

The extended analysis in this paper seeks to explore other predictive variables and methods from machine learning, to interpret their effectiveness on a limited dataset. In the previous section, splitting the patients into an older and younger cohort showed a difference in probability of survival. However, the distribution of age in the survivor and non-survivor groups as shown in Figure 10 has a large overlap and no significant difference between the two groups split at 60 years old (black dashed line).

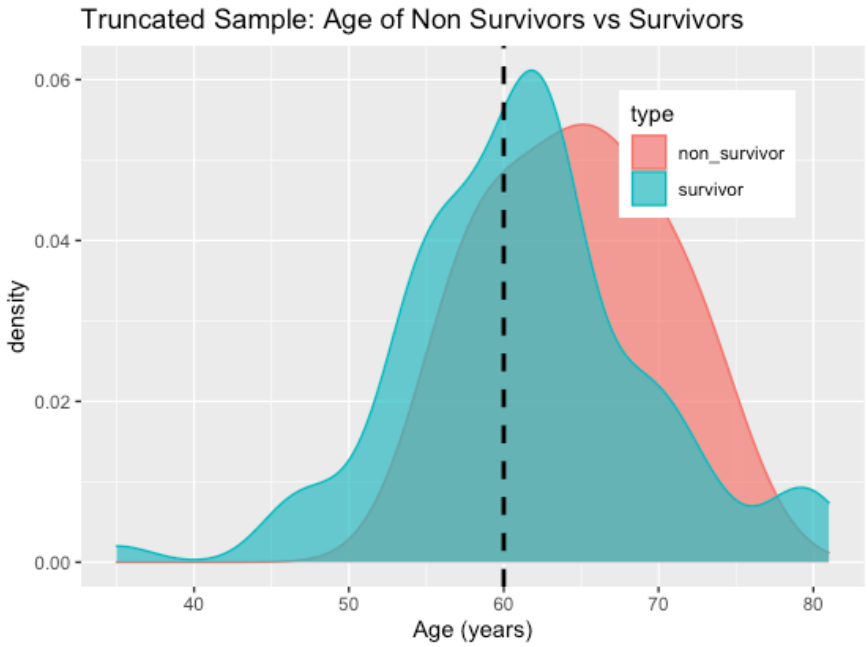

Figure 10. Distribution of age for non survivors and survivors

A correlation matrix looking at other possible predictors is displayed in Figure 11 and shows the correlation coefficients between pairs of variables. A value of 1 indicates a positively correlated relationship, a value of 0 indicates no relationship, and a value of -1 indicates a negatively correlated relationship. The colors of the squares indicates the strength of the relationship corresponding with the gradient legend to the right of the plot. The strongest negative correlation with survival time is epss (E-point septal separation), a measure of contractility of the heart. Myocardial contractility is the ability of the heart muscle to contract. A higher epss means more abnormal motion, and thus a lower survival period is expected. The strongest positive correlation with survival time is fractional shortening, another measure of contractility. Lower numbers of fractional shortening indicate increasingly abnormality. Therefore, a lower survival period is expected for decreasing fractional shortening values. Age , pericardial effusion, fractional shortening, epss, lvdd, and wall motion index are used to predict if the one year survival of the patient in the algorithms.

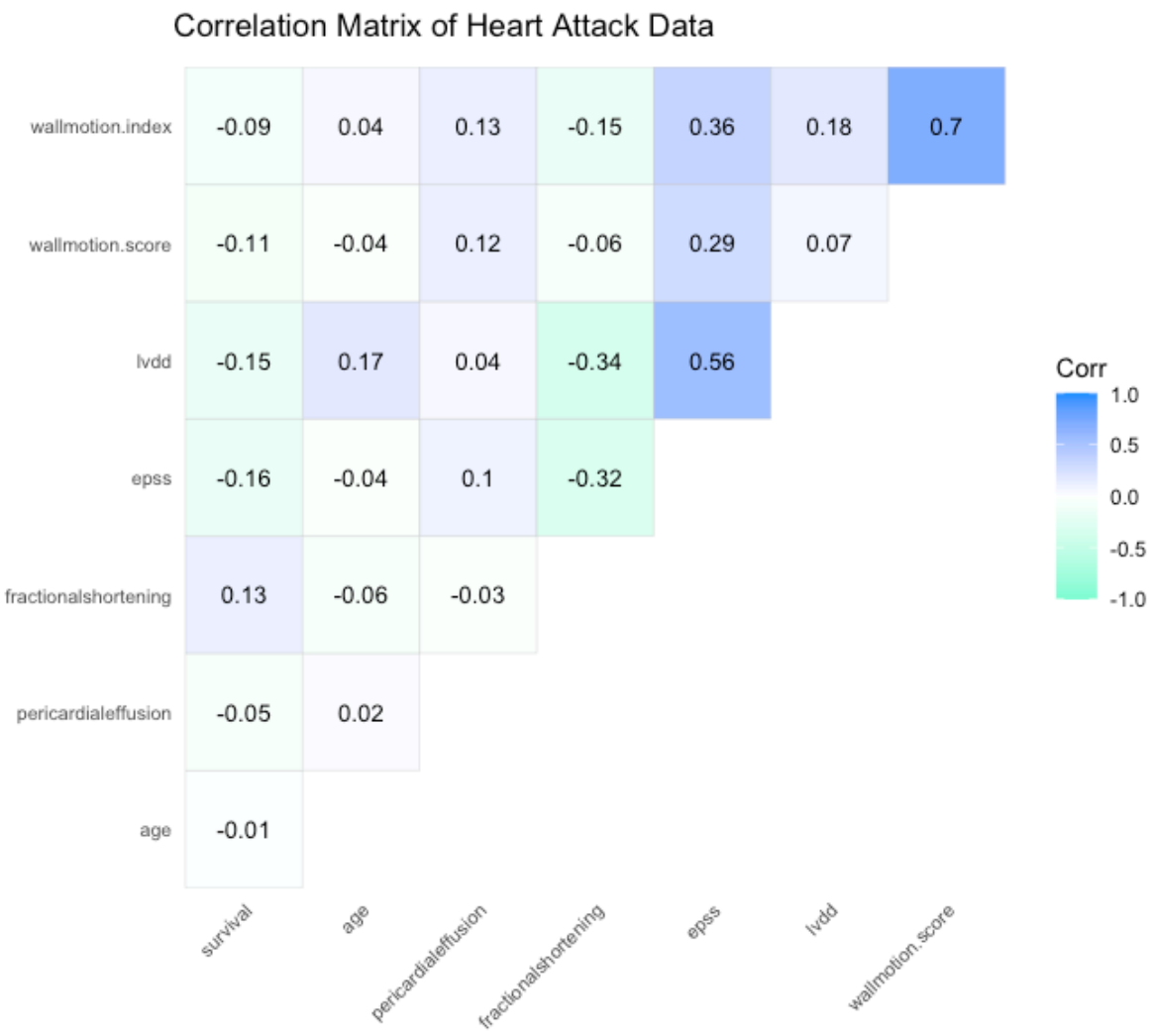

Figure 11. Correlation matrix of variables in available data

Logistic Regression

A logistic regression is applied to determine the probability of a patient being a survivor or non-survivor. Age , pericardial effusion, fractional shortening, epss, lvdd, and wall motion index are the variables used in the regression. The fitted regression coefficients on the training dataset and their corresponding p-values are provided in Table 3. None of the variables are particularly significant at predicting the probability since their p-values are quite large and exceed the 5% significance level. We cannot reject the null hypothesis that they have no effect on the predictor. The accuracy of the model predictions on the test data is only 62%, albeit better than using WMS threshold to predict alone.

Table 3. Logistic Regression Coefficients

	Estimate	Std. Error	p-value
(Intercept)	0.53	3.33	0.87
age	0.03	0.04	0.48
pericardialeffusion1	0.06	0.63	0.93
Fractional shortening	-5.24	3.74	0.16
epss	0.00	0.07	0.97
lvdd	-0.11	0.39	0.79
Wall motion index	-0.48	0.89	0.59

KNN-K Nearest Neighbors

A KNN analysis using age , pericardial effusion, fractional shortening, epss, lvdd, and wall motion index was used to predict whether a patient is a survivor or non-survivor based on the labels of its nearest neighbors (or observations most 'similar' to it). As shown in Figure 12, the number of neighbors used in this algorithm is 7, based on the optimal accuracy in cross-validation. The accuracy of the predictions in this model is 48%, better than using WMS alone but worse than logistic regression. This is possibly due to the lack of unique data points for the non-survivors group, which makes it hard to determine good predictive neighbors.

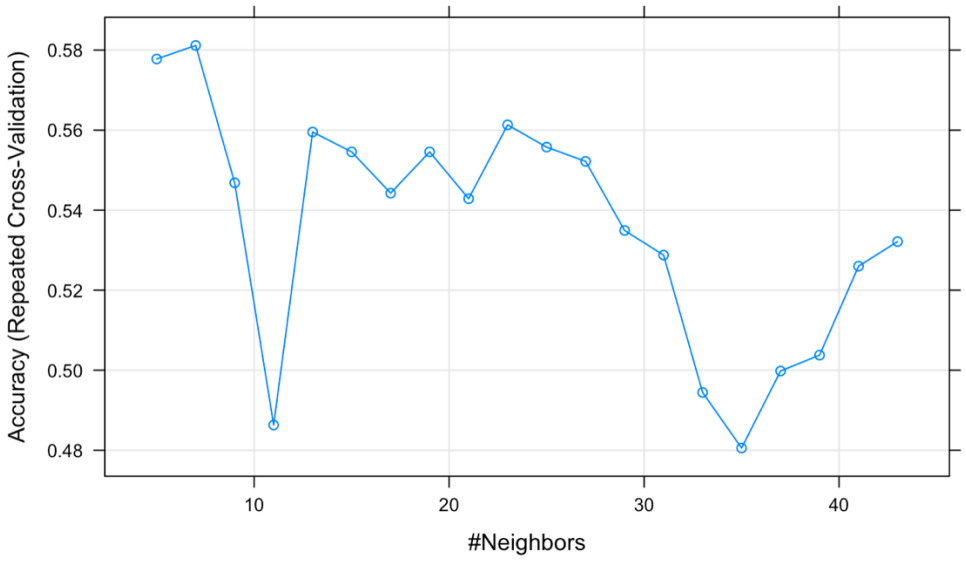

Figure 12. Accuracy vs Number of Neighbors in KNN Analysis

Random Forest

A random forest algorithm on the same variables as in the other two models was run. The random forest is made up of the aggregate predictions of many decision tree models. The algorithm determined the relative importance of each variable in the prediction, as shown in Figure 13. The most important variable was epss while the least important was presence of pericardial effusion. Wall motion index ranked low on the scale of importance; this indicates that it is not a good variable for prediction for this dataset as was shown in previous sections. The accuracy of the random forest model was 90%, better than both logistic regression and KNN.

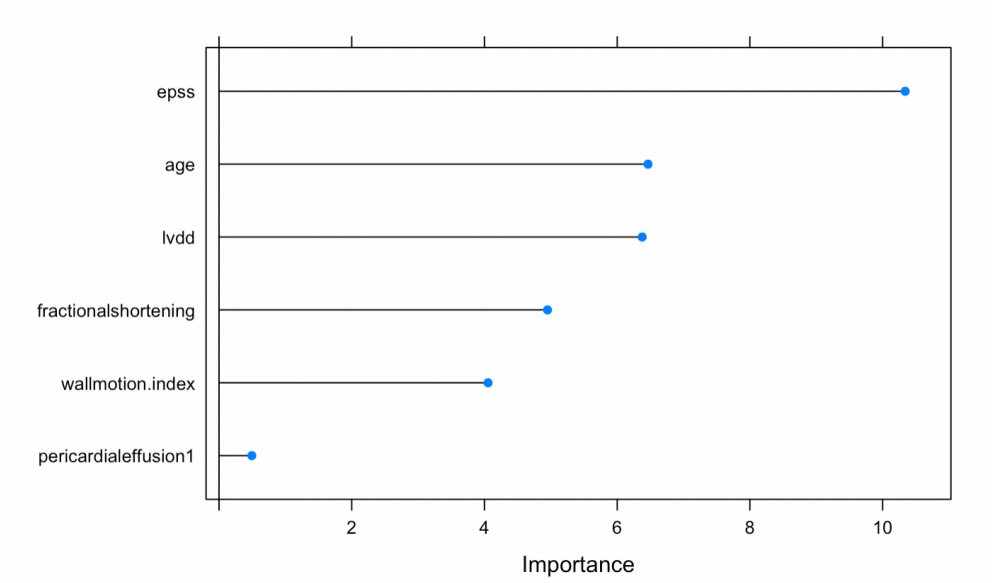

Figure 13. Importance of Variables in Random Forest Analysis

Comparison of Algorithms

The Receiver Operating Characteristic (ROC) curve for all three models is shown in Figure 14. A ROC curve displays the predictive ability of a classification model as the threshold for classification is varied. The area under the curve (AUC) is a measure of the predictive accuracy of the model. The AUC is the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. The straight black line on the curve is the probability of picking by chance, in other words anything below it has no predictive value. The bigger the AUC, the better is model is in general. From the plot, it is clear that the Random Forest model is the best one for prediction for this dataset. This is also reflected in the accuracy, sensitivity and specificity values for each model compared in Table 4. KNN is the worst model to use and has low predictive value, likely due to the lack of unique observations for the non survivors group.

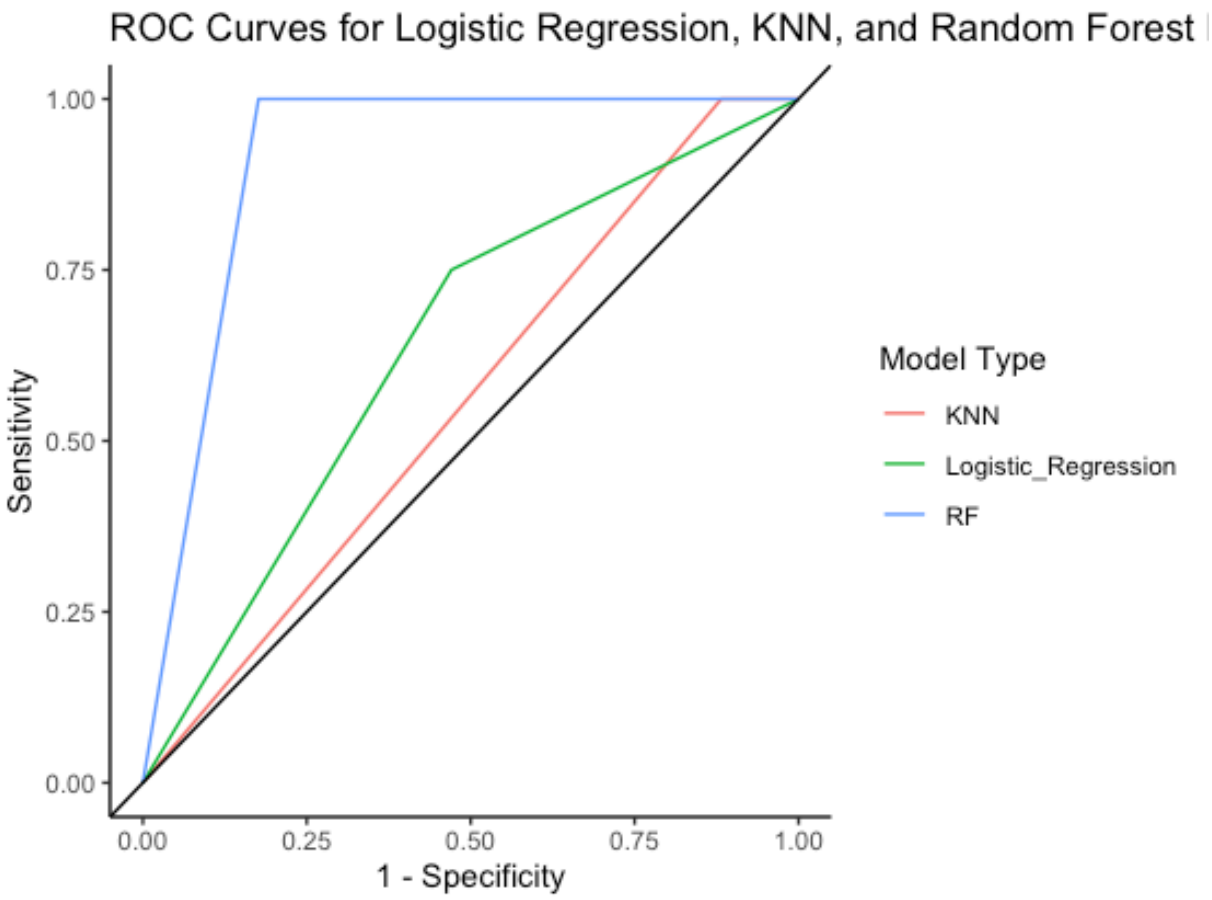

Figure 14. ROC Curves for Logistic Regression, KNN, and Random Forest Analyses

Table 4. Model Evaluation measures

	Accuracy	**Sensitivity**	**Specificity (TPR)**	**Negative predictive value (TNR)**
Logistic Regression	62%	53%	75%	53%
KNN	48%	11%	100%	44%
Random Forest	90%	82%	100%	80%

Conclusion

This paper is a comparative analysis of the study undertaken by Kan et al. (1986) in "Short and long term predictive value of admission wall motion score in acute myocardial infarction". The original paper analyzed hospital echocardiogram data for a sample of 345 patients diagnosed with myocardial infarction (heart attack). The data analyzed in this paper is a truncated version of the full data, with 132 patient observations (96 observations without censored data). In the original study, the authors determined a wall motion score (WMS) of 10 is the best threshold for predicting whether a patient will die within a year. The wall motion score distributions of the survivor and non-survivor groups were compared in this analysis and determined to be unrepresentative of the original sample, thus the exact results could not be replicated using the same WMS predictor.

An extended analysis used Kaplan-Meier estimators to predict the probability of survival over time, which also preserved the censored data. It was found that a threshold WMS score of 13 is a better predictor for survival for this particular sample. However, a t-test for the sample means found no significant difference in wall motion scores of the survivor and non-survivor groups.

The new analysis in this paper also evaluated other variables in the data and their correlations, as well as supervised machine learning algorithms. Random resampling (bootstrap method) was used to balance the samples and create a final sample of 50% survivors -50% non-survivors to train and test the algorithms. Logistic regression, K Nearest Neighbors (KNN), and random forest algorithms were applied using the following features: age , pericardial effusion, fractional shortening, E point septal separation (epss), left ventricular end-diastolic dimension (lvdd), and wall motion index (a standardized form of wall motion score). The best model is the random forest model, with an accuracy of 90%, sensitivity of 82%, and specificity of 100%. The worst model was KNN, with accuracy of 48%. The random forest model showed that the most important variable in prediction is epss, while wall motion index ranked second to last in importance. This indicates that at least for this particular dataset, wall motion index and score are not the critical variables for predicting survival.

The limitations in this work stemmed from the small dataset. First the dataset is a subset of the full dataset in the original paper; only 28% of comparable observations compared to the original. Second, the dataset is not representative of the original data in terms of the distribution of the variables, especially WMS. Third there is an imbalance in the groups which makes modeling and prediction difficult; the data is heavily skewed with 92% survivors and 8% non-survivors. Therefore, while the results of this analysis does not agree with the original paper results, it does not mean the original results were disproven. Future work would involve obtaining the original full dataset and replicating the analysis with those sample to obtain a comparable evaluation. As other studies in the field have been conducted after the original paper, it will be helpful to analyze their findings and improve upon the original. Another consideration is increasing the study size to an even larger sample, in order to find more discernable trends. With these recommendations in place, a more informative analysis and results may be reached that will conclusively support or challenge the original work by Kan et al.

Bibliography

Benjamin, EJ., Muntner, P., Alonso, A., Bittencourt. MS., Callaway, CW., Carson, AP., et al. (2019). Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation. 2019;139(10):e56–528.

Kan, G, Visser, C A, Koolen, J J, & Dunning, A J. (1986). Short and long term predictive value of admission wall motion score in acute myocardial infarction. British Heart Journal, 56(5), 422–427. https://doi.org/10.1136/hrt.56.5.422

Klein, Patrick, Holman, Eduard R, Versteegh, Michel I.M, Boersma, Eric, Verwey, Harriette F, Bax, Jeroen J, Dion, Robert A.E, & Klautz, Robert J.M. (2009). Wall motion score index predicts mortality and functional result after surgical ventricular restoration for advanced ischemic heart failure. European Journal of Cardio-Thoracic Surgery, 35(5), 847–853. https://doi.org/10.1016/j.ejcts.2008.12.046

Torio, Celeste (August 2013). "National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011". HCUP.