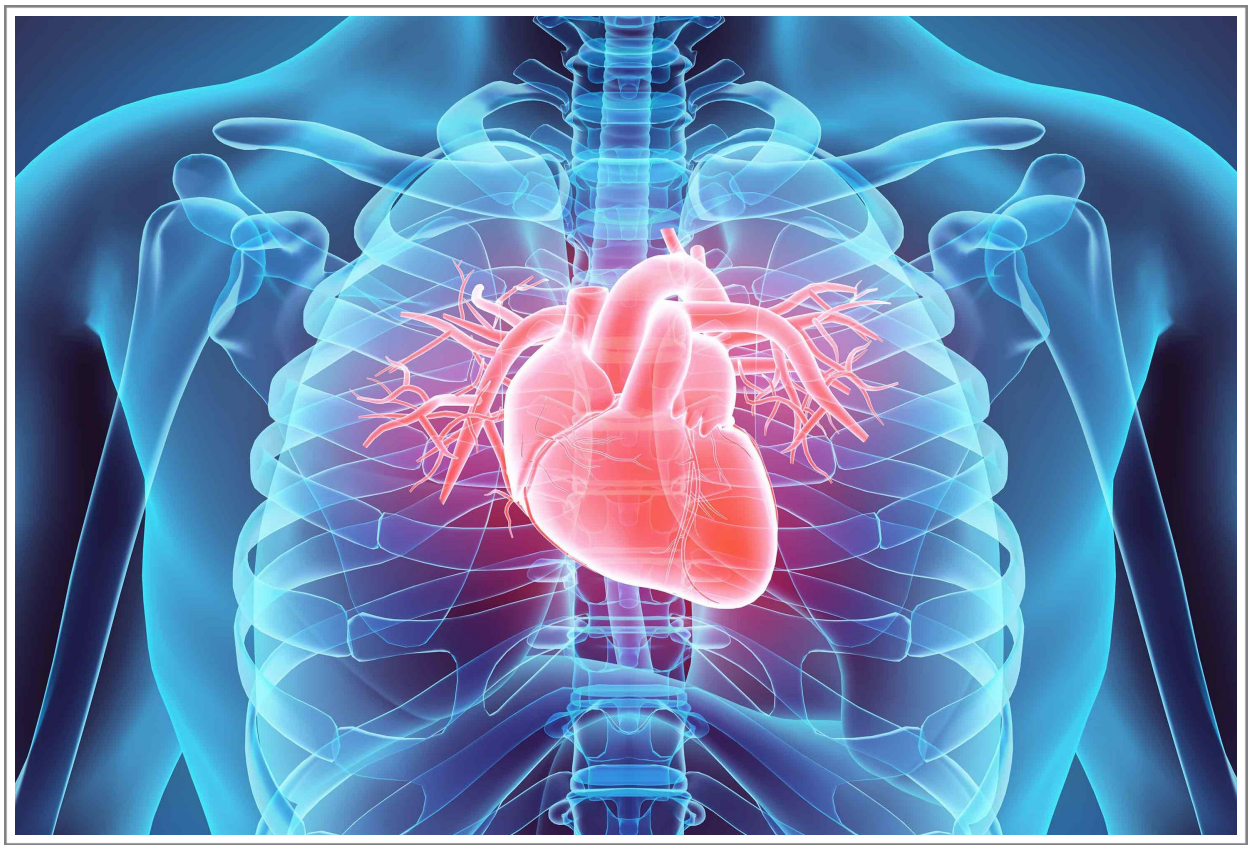
Cardiovascular Disease Analysis

PREDICTING DEATH OCCURRENCE FROM PATIENT DATA



JAE KYOUNG LEE

Introduction

Cardiovascular diseases are responsible for the death of 17 million people annually. Types of cardiovascular diseases include coronary heart disease, heart failure, cerebral-vascular disease and many more. Predicting deaths with high accuracy and sensitivity is a challenge, but essential for physicians to keep mortality rates low.

The data we will be using in this analysis is the Heart Failure Clinical Records data set found in the infamous University of California Irvine Machine Learning repository[1] . The data set contains medical records of 299 heart failure patients collected at two hospitals in Punjab, Pakistan between April and December 2015. The original dataset version was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) and made available by them on FigShare under the Attribution 4.0 International (CC BY 4.0: freedom to share and adapt the material) copyright in July 2017. Some questions we want to explore will include the following:

1. How is the data different between death and no-death occurrences?
2. What variables may cause death(among predictors in data)?
3. What models can predict death occurrence with high accuracy?
4. Do follow-up periods effect death occurrence?
5. Is our final model usable in practice(High accuracy, sensitivity, specificity)?

[1] https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records : Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020). [Web Link]

Data Description

Thirteen (13) clinical features:

Categorical variables:

- anaemia: decrease of red blood cells or hemoglobin (boolean)(categorical)
- high blood pressure: if the patient has hypertension (boolean)(categorical)
- diabetes: if the patient has diabetes (boolean)(categorical)
- sex: woman or man (binary)(categorical)
- smoking: if the patient smokes or not (boolean)(categorical)

Continuous variables:

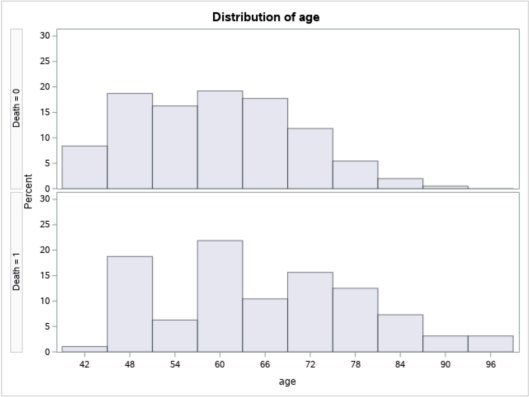
- age: age of the patient (years)(continuous)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)(continuous)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)(continuous)
- platelets: platelets in the blood (kiloplatelets/mL)(continuous)
- serum creatinine: level of serum creatinine in the blood (mg/dL)(continuous)
- serum sodium: level of serum sodium in the blood (mEq/L)(continuous)
- time: follow-up period (days)(continuous)

Response variable:

- [target] death event: if the patient deceased during the follow-up period (boolean)

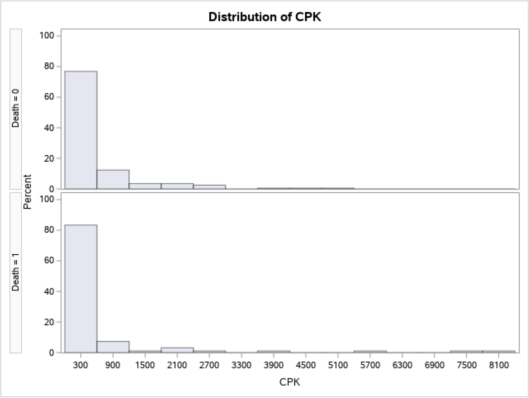
Analysis

Beginning the analysis, we want to have general descriptive overview of medical records characteristics for the patients considered in this study to better understand the data. We want to know how continuous data is distributed when classified by death occurrence.



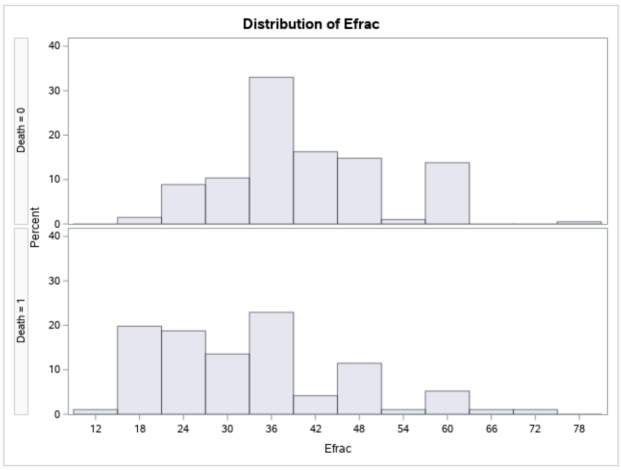
<Figure 1-1> Distribution of age classified by death occurrence

Starting with age, we can see that the age distribution among death and no death occurrence is quite different(Figure 1-1). For no death occurrence, the data is highly distributed with age 40 to 70(mean = 58.76, sd = 10.64). For death occurrence, the data is highly distributed with ages 48 to 80(mean = 65.22, sd = 14.22). (See appendix for additional descriptive statistics(Table 1)).



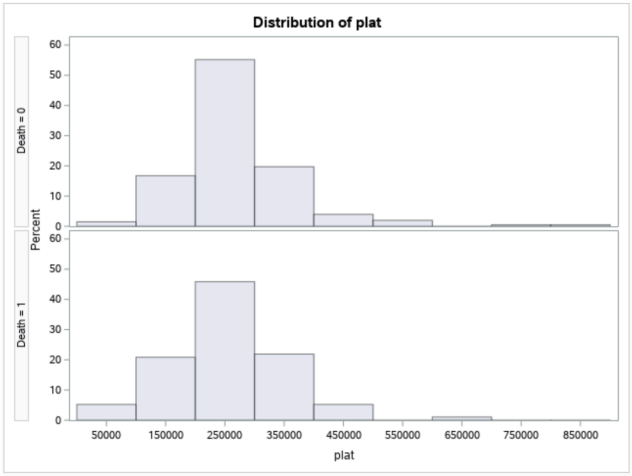
<Figure 1-2> Distribution of creatinine phosphokinase(CPK) classified by death occurrence

Above we see the distributions for CPK levels(mcg/L)(Figure 1-2). The data is distributed nearly identically among death(mean = 540.0542, sd = 753.8) and no death occurrences(mean = 670.2, sd = 1317).



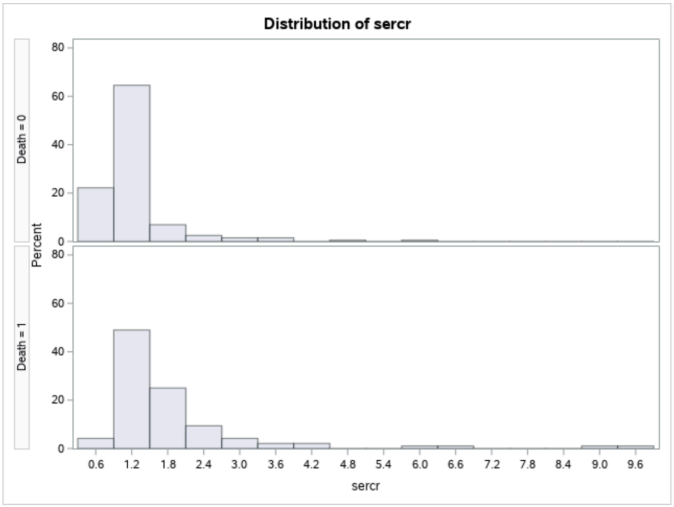
<Figure 1-3> Distribution of ejection fraction(Efrac) classified by death occurrence

The third variable we look at is Efrac(Ejection Fraction(%)), and the distributions among death and no death occurrence is different(Figure 1-3). For no death occurrence, the data is highly distributed with Efrac percentages of 30 to 50(mean = 40.26, sd = 10.86). For death occurrence, the data is highly distributed with percentages of 18 to 48(mean = 33.47, sd = 12.53).



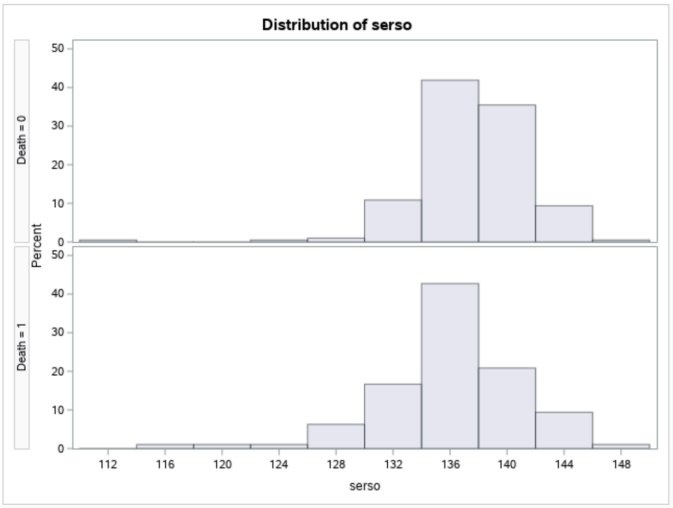
<Figure 1-4> Distribution of platelets(plat) classified by death occurrence

The fourth continuous variable is plat(platelets(kiloplatelets/L)), and the distributions among death(mean = 266657.5, sd = 97531) and no death occurrence(mean = 256381, sd = 98526) are almost identical(Figure 1-4).



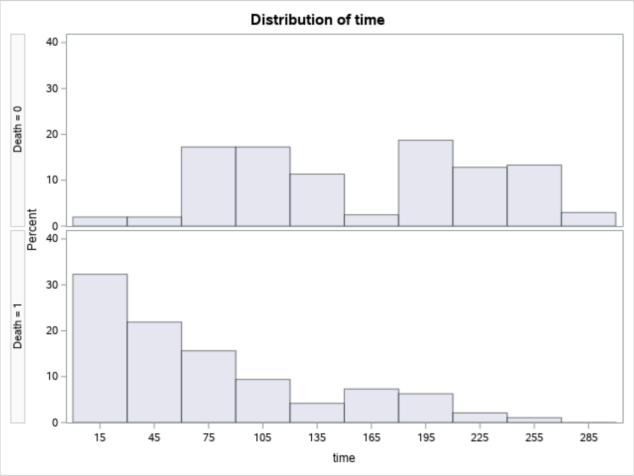
<Figure 1-5> Distribution of serum creatinine classified by death occurrence

The fifth variable is sercr(serum creatinine(mg/dL)), and the distributions among death(mean = 1.184877, sd = 0.65) and no death occurrence(mean = 1.835833, sd = 1.47) are quite different. The mean of sercr in no death occurrence is nearly twice the mean in death occurrence(Figure 1-5).



<Figure 1-6> Distribution of serum sodium classified by death occurrence

The sixth variable is serso(serum sodium(mEq/L)), and the distributions among death(mean = 137.2167, sd = 3.98) and no death occurrence(mean = 135.375, sd = 5.001) are almost identical(Figure 1-5).



<Figure 1-7> Distribution of follow-up period(time) classified by death occurrence

The last continuous variable time(days) show an unequal data distribution among death and no death occurrences(Figure 1-7). For no death occurrence, the data is almost evenly distributed with days 55 to 275(mean = 158.3399, sd = 67.74), while for death occurrence the data is highly distributed with days 15 to 105(mean = 70.88542, sd = 62.38).

Table of Death by anae(Figure 2-1)

Death / Frequency / Expected	anae 0	anae 1	Total
1	50	46	96
	54.582	41.418	
0	120	83	203
	115.42	87.582	
Total	170	129	299

Table of Death by diab(Figure 2-2)

Death / Frequency / Expected	diab 0	diab 1	Total
1	56	40	96
	55.866	40.134	
0	118	85	203
	118.13	84.866	
Total	174	125	299

Table of Death by hbp(Figure 2-3)

Death / Frequency / Expected	hbp 1	hbp 0	Total
1	39	57	96
	33.712	62.288	
0	66	137	203
	71.288	131.71	
Total	105	194	299

Table of Death by sex(Figure 2-4)

Death / Frequency / Expected	sex 1	sex 0	Total
1	62	34	96
	62.288	33.712	
0	132	71	203
	131.71	71.288	
Total	194	105	299

Table of Death by smk(Figure 2-5)

Death / Frequency / Expected	smk 0	smk 1	Total
1	66	30	96
	65.177	30.823	
0	137	66	203
	137.82	65.177	
Total	203	96	299

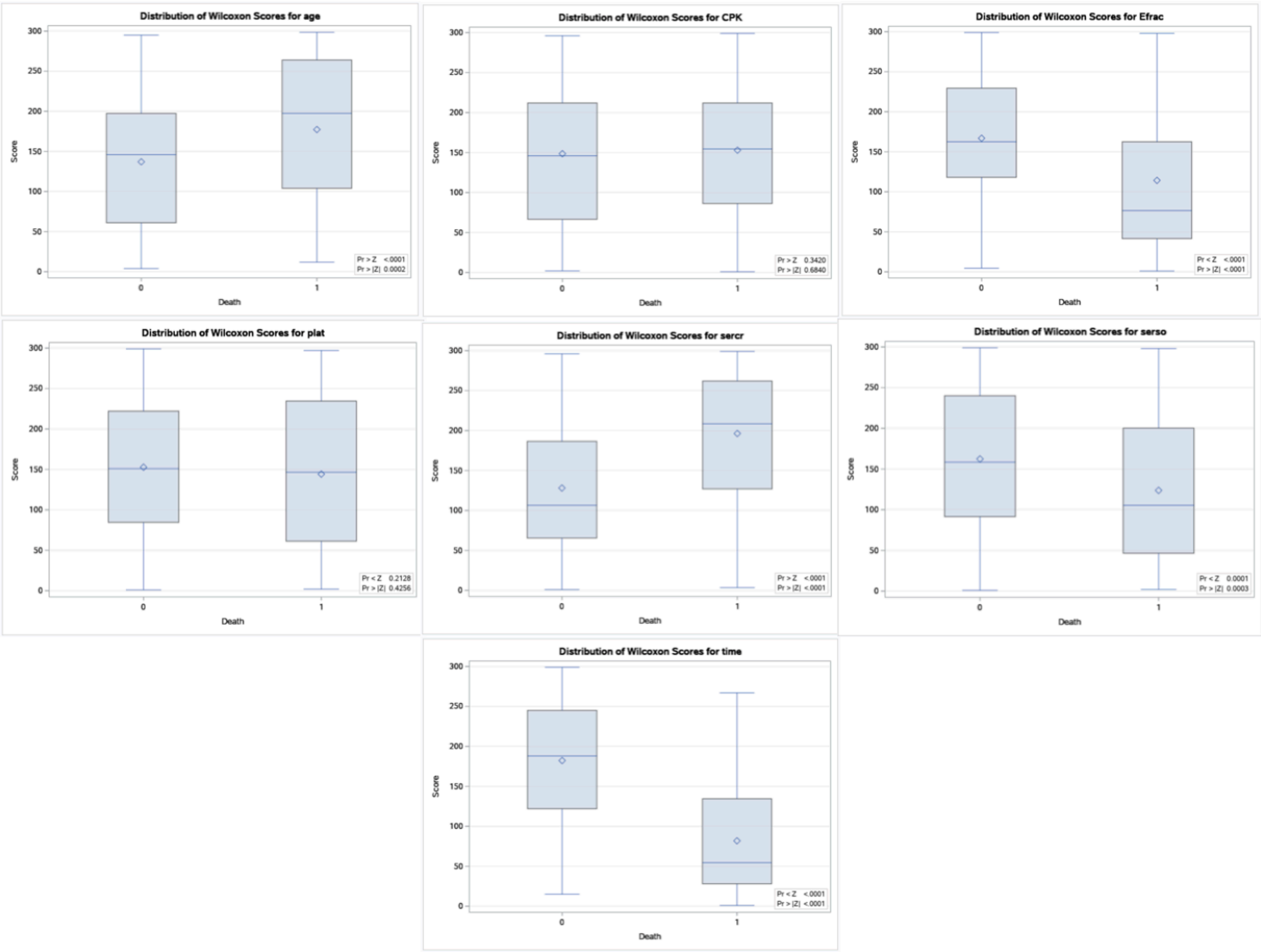
<Figure 2-1, 2-2, 2-3, 2-4, 2-5> Contingency tables for binary variables anae, diab, hbp, sex, smk

For categorical(binary) variables, we created contingency tables for each variable by death occurrence to see how much frequency counts and expected counts are deviated(Fig 2-1, 2-2, 2-3, 2-4, 2-5). For variables anae(anaemia), diab(diabetes), hbp(high blood pressure), sex, and smk(smoking status), we can see that the frequency counts are not deviated much from the

expected counts. This may be a good indication that there is no association between death and these variables.

Now we move on to perform a statistical analysis on the possible significant association between death occurrence and each of the binary variables(anae, diab, hbp, sex, smk). Since the expected cell counts are relatively large for all the binary variables, we conducted the Pearson's Chi-Square test of independence for each. The null hypothesis is that there is no association between the variables and the observed counts are not different from the ones expected under an independence assumption. Results indicate that we fail to reject the null hypothesis for all of the binary variables($p > 0.05$)(Table 3-1 ~ 3-5 in Appendix). The statistical analysis on the possible significant association between death occurrence and each of the binary variables(anae, diab, hbp, sex, smk) shows that we cannot say any of these variables shows association with death occurrence. This is quite surprising given the fact that most of the variables especially smoking is known to be extremely harmful to a patient. This analysis would most likely improve if we could collect ordinal data assigned to several levels of use(smoking) or severity(anae, diab).

Now that we have performed hypothesis tests on the binary variables, we will move on to a statistical analysis comparing the mean values of the continuous predictors(age, CPK, Efrac, plat, sercr, serso, time) between death and no-death occurrences. Before we do this, it is important that we check for normality of each predictors with tests for normality. If normality is unreasonable, we must use non-parametric tests instead of t-test. All four tests indicate that we must reject the null hypothesis that the variables are normally distributed($p < 0.05$)(Table 4-1 ~ 4.7 in Appendix). Therefore we must use the Wilcoxon test to compare the mean values of the continuous predictors between death and no-death occurrences.



<Figure 5-1 ~ 5-7> Distribution of Wilcoxon Scores for continuous variables

Before we address the results of the Wilcoxon test, we can see above that most of the variables have different mean scores. Variables age(Figure 5-1), Efrac(Figure 5-3), sercr(Figure 5-5), serso(Figure 5-6) and time(Figure 5-7) especially stand out to the eye. Since it is difficult to assess the significance by solely looking at these plots, we must look at the results of the Wilcoxon test that does not assume normality(Table 5-1 ~ 5-7 in Appendix). The test results tells us that variables age, Efrac, sercr, serso, and time have significantly different means between death occurrence and no death occurrence. On the other hand, variables CPK and plat do not have significantly different means between death occurrence and no death occurrence.

Modeling

Now we fit a binary logit model to the binary variable Death as a function of the remaining variables excluding the variable time(Table 6 in the Appendix). We exclude time because we want to try to fit a conditional logistic regression with the stratification time later. Using the stepwise selection method to choose the best predictors for our model, variables age, Efrac, and sercr were significant(p-value = < 0.0001). The model tells us that 1 year increase in age would have a multiplicative effect of 1.053 on the odds of death occurrence, 1% increase in ejection fraction would have a multiplicative effect of 0.932(which means it will decrease), and 1 mg/dL increase of serum creatinine would have a multiplicative effect of 1.946. The criterion for assessing whether the model has improved will be a metric called Akaike information criterion(AIC). The AIC of this model is 377.349. We will see if we can improve AIC with other methods.(See Table 6 in Appendix for detailed model information). Below we can see the frequency table to compare the levels observed(Death = 0 or 1)(using cutoff 0.4 for classifying into Death =1).

Table of Death by Predictions

	Predictions		
Death	**0**	**1**	**Total**
0	172	31	203
1	34	62	96
Total	206	93	299

<Table 8> Frequency table of logit model with predictors age, Efrac, sercr

Looking at the frequency table of response Death by predictions(Table 8), sensitivity is 64.6%, specificity is 84.7%, and overall accuracy is 78.26%.

As we mentioned before, the follow-up time variable has a different extent for each patient. it might be important to consider the different follow-up periods and stratify the patients according to the different follow-up period lengths. The follow-up period is going to be mapped into a new variable called month. If time < 30, month = 0. If time > 60, month = 2. Otherwise month = 1. We will now fit a conditional logistic regression with the stratification “month”(Table 7). Again

using the stepwise selection method to choose the best predictors for our model, variables age, Efrac, and sercr were deemed significant($p < 0.05$). The model tells us that 1 year increase in age would have a multiplicative effect of 1.052 on the odds of death occurrence, 1% increase in ejection fraction would have a multiplicative effect of 0.923(which means it will decrease), and 1 mg/dL increase of serum creatinine would have a multiplicative effect of 1.938. The AIC of this model is 260.375, which is better than the logit model before.(See Table 7 in Appendix for detailed model information). Below we can see the frequency table to compare the levels observed(Death = 0 or 1)(using cutoff 0.4 for classifying into Death =1).

Table of Death by Predictions

	Predictions		
Death	**0**	**1**	**Total**
0	186	17	203
1	26	70	96
Total	212	87	299

<Table 9> Frequency table of conditional logit model with predictors age, Efrac, sercr(stratification with month)

Looking at the frequency table of response Death by predictions(Table 9), sensitivity increased to 72.91%, specificity increased to 91.6%, and overall accuracy also increases to 85.61%. One concern is that sensitivity is too low. Sensitivity is arguably more important than specificity in medical situations because a high rate of false negatives(27.09 in this case)may lead to bad circumstances.

Looking at the balance of the target variable Death, we have 203 0's(no death occurrence) and 96 1's(death occurrence). While this is not a huge imbalance, it may be the reason why a 26 out of 42 misclassifications are false negatives. To address this issue, we may use the oversampling method of duplicating examples in the minority class so that we can make the dataset exactly balanced(Figure 6 in Appendix). With this new data with 406 samples, we fit a conditional logistic regression(Table 10 in Appendix for model information). The model tells us that 1 year increase in age would have a multiplicative effect of 1.065 on the odds of death occurrence, 1% increase in ejection fraction would have a multiplicative effect of 0.908(which means it will decrease), and 1 mg/dL increase of serum creatinine would have a multiplicative effect of 2.6. Although the AIC of this model is increased to 400.44, we will not use this metric to compare the

models before since we have a different dataset. Below we can see the frequency table to compare the levels observed(Death = 0 or 1)(using cutoff 0.4 for classifying into Death =1).

Table of Death by Predictions

	Predictions		
Death	**0**	**1**	**Total**
0	170	33	203
1	16	187	203
Total	186	220	406

<Table 11> Frequency table of conditional logit model with predictors age, Efrac, sercr(stratification with month) **after oversampling method**

Looking at the frequency table of response Death by predictions(Table 11), sensitivity increased significantly to 92.1%, specificity decreased to 83.7%, and overall accuracy increases to 87.93%. This means that our oversampling method improved our results by achieving a lower false negative rate and higher overall higher accuracy. Since we have duplicated some data, this may introduce some bias to the results. What we can learn from this analysis is that we should collect nearly equal amounts of data from each class(in our case, Death = 0 or 1). Also, it will be important that we collect accurate data for variables age, ejection fraction, serum creatinine, and month. Expanding and balancing the dataset will also be a big improvement for future analyses.

Conclusions

Summary of Analysis:

In Figure 1-1 through Figure 1-7(page 4 ~ 7), we saw the distributions of each continuous predictor between death and no death occurrences. We realize that all the significant predictors in our conditional logistic regression had significantly different distributions between the two classes(age, Efrac, sercr, time(later mapped into variable month)). For categorical variables, we created contingency tables to see whether there were significant deviations between expected frequency and frequency(Fig 2-1 ~ 2-5, page 7), and we learned that all of the categorical variables did not. This directly leads to our results for the Pearson's Chi-square tests that indicated all the categorical variables failed to having a significant association with the target response(Death)($p > 0.05$). To further test whether our continuous variables had a significant difference in mean values between death and no death occurrences, we performed the Wilcoxon test. Results indicated that variables age, Efrac, sercr, serso, and time had significantly different wilcoxon scores. After exploring the data and performing statistical analysis on the data, we moved on to fit our first logistic regression. Using the stepwise selection method to choose signifiant predictors, the model selected predictors age, Efrac(ejection fraction), and sercr(serum creatinine). Sensitivity was 64.6%, specificity was 84.7%, and overall accuracy was 78.26% (AIC = 377.349). Not bad for a first, but we had to improve this. The second model we fit is a conditional logistic regression with the stratification month(1, 2 or 3), our new variable we constructed from the variable time. This led to better results, with sensitivity increasing to 72.91%, specificity increasing to 91.6%, and overall accuracy also increasing to 85.61%. The final model we fit is the same conditional logistic regression but with the data engineered with the oversampling method. This resulted in sensitivity increasing significantly to 92.1%, specificity decreasing to 83.7%, and overall accuracy increasing to 87.9%.

Overall conclusion for Physicians:

Results of the statistical analyses that we performed indicate that all the categorical variables including anaemia, high blood pressure, creatinine phosphokinase, diabetes, and smoking are not significant for predicting death occurrence($p > 0.05$). Although this does not mean that they are completely independent, including them in our logistic models would not improve results. If these variables were collected to be continuous rather than binary, it will most likely result in improving our analysis. On the other hand, continuous variables age, ejection fraction, serum creatinine, and follow-up periods proved to be extremely significant for predicting death

occurrence($p < 0.05$). Collecting precise results of these variables will be vital for future analyses. Physicians should be aware of the following results obtained from our conditional logistic regression:

Strata Summary			
	Death		
Month	1	0	Frequency
<1 month	45	2	47
1 month	57	6	63
>2 months	101	195	296

We can see that when follow-up period was less than 1 month, 45 out of 47 patients deceased. The rate increases each month with 57 out of 63 deaths in more than 1 month, and 101 out of 296 in more than 2 months. After inspecting this table, we found out that 33 out of 43 observations that were misclassified had follow-up periods with larger than 2 months. Although we attempted to find new patterns to variable month by classifying time further to 9 months, there was no significant pattern after month 2. Since sensitivity and accuracy is high with 93.1% and 87.2% respectively for our final model, it is definitely usable in practice for assisting physicians in predicting death occurrences of patients. To prevent low mortality, it will be essential to keep ejection fraction between the normal range of 50 to 70%, and serum creatine levels between the normal range of 0.9 to 1.3.(See Table 1 for detailed descriptive statistics).

Appendix

<Table 1> Descriptive statistics of continuous variables when classified by death occurrence

Basic Statistical Measures(age) Death = 0			
Location		Variability	
Mean	58.76191	Std Deviation	10.63789
Median	60.00000	Variance	113.16471
Mode	60.00000	Range	50.00000
		Interquartile Range	15.00000

Basic Statistical Measures(age) Death = 1			
Location		Variability	
Mean	65.21528	Std Deviation	13.21456
Median	65.00000	Variance	174.62448
Mode	60.00000	Range	53.00000
		Interquartile Range	20.00000

Basic Statistical Measures(CPK) Death = 0			
Location		Variability	
Mean	540.0542	Std Deviation	753.79957
Median	245.0000	Variance	568214
Mode	582.0000	Range	5179
		Interquartile Range	473.00000

Basic Statistical Measures(CPK) Death = 1			
Location		Variability	
Mean	670.1979	Std Deviation	1317
Median	259.0000	Variance	1733385
Mode	582.0000	Range	7838
		Interquartile Range	453.50000

Basic Statistical Measures(Efrac) Death = 0			
Location		Variability	
Mean	40.26601	Std Deviation	10.85996
Median	38.00000	Variance	117.93879
Mode	35.00000	Range	63.00000
		Interquartile Range	10.00000

Basic Statistical Measures(Efrac) Death = 1			
Location		Variability	
Mean	33.46875	Std Deviation	12.52530
Median	30.00000	Variance	156.88322
Mode	25.00000	Range	56.00000
		Interquartile Range	13.00000

Basic Statistical Measures(plat) Death = 0			
Location		Variability	
Mean	266657.5	Std Deviation	97531
Median	263000.0	Variance	9512335419
Mode	263358.0	Range	824900
		Interquartile Range	83000

Basic Statistical Measures(plat) Death = 1			
Location		Variability	
Mean	256381.0	Std Deviation	98526
Median	258500.0	Variance	9707310182
Mode	263358.0	Range	574000
		Interquartile Range	115000

Basic Statistical Measures(sercr) Death = 0			
Location		Variability	
Mean	1.184877	Std Deviation	0.65408
Median	1.000000	Variance	0.42782
Mode	1.000000	Range	5.60000
		Interquartile Range	0.30000

Basic Statistical Measures(sercr) Death = 1			
Location		Variability	
Mean	1.835833	Std Deviation	1.46856
Median	1.300000	Variance	2.15667
Mode	1.000000	Range	8.80000
		Interquartile Range	0.85000

Basic Statistical Measures(serso) Death = 0			
Location		Variability	
Mean	137.2167	Std Deviation	3.98292
Median	137.0000	Variance	15.86368
Mode	137.0000	Range	35.00000
		Interquartile Range	5.00000

Basic Statistical Measures(serso) Death = 1			
Location		Variability	
Mean	135.3750	Std Deviation	5.00158
Median	135.5000	Variance	25.01579
Mode	134.0000	Range	30.00000
		Interquartile Range	5.50000

Basic Statistical Measures(time) Death = 0			
Location		Variability	
Mean	158.3399	Std Deviation	67.74287
Median	172.0000	Variance	4589
Mode	187.0000	Range	273.00000
		Interquartile Range	118.00000

Basic Statistical Measures(time) Death = 1			
Location		Variability	
Mean	70.88542	Std Deviation	62.37828
Median	44.50000	Variance	3891
Mode	10.00000	Range	237.00000
		Interquartile Range	79.50000

<Table 3-1 ~ 3-5> Pearson Chi-Square tests for the assumption of independence(Binary variables)

Statistics for Table of Death by anae(Table 3-1)

Statistic	DF	Value	Prob
Chi-Square	1	1.3131	0.2518
Likelihood Ratio Chi-Square	1	1.3086	0.2527
Continuity Adj. Chi-Square	1	1.0422	0.3073
Mantel-Haenszel Chi-Square	1	1.3087	0.2526
Phi Coefficient		-0.0663	
Contingency Coefficient		0.0661	
Cramer's V		-0.0663	

Statistics for Table of Death by diab(Table 3-2)

Statistic	DF	Value	Prob
Chi-Square	1	0.0011	0.9732
Likelihood Ratio Chi-Square	1	0.0011	0.9732
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0011	0.9732
Phi Coefficient		0.0019	
Contingency Coefficient		0.0019	
Cramer's V		0.0019	

Statistics for Table of Death by hbp(Table 3-3)

Statistic	DF	Value	Prob
Chi-Square	1	1.8827	0.1700
Likelihood Ratio Chi-Square	1	1.8630	0.1723
Continuity Adj. Chi-Square	1	1.5435	0.2141
Mantel-Haenszel Chi-Square	1	1.8764	0.1707
Phi Coefficient		0.0794	
Contingency Coefficient		0.0791	
Cramer's V		0.0794	

Statistics for Table of Death by sex(Table 3-4)

Statistic	DF	Value	Prob
Chi-Square	1	0.0056	0.9405
Likelihood Ratio Chi-Square	1	0.0056	0.9405
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0056	0.9406
Phi Coefficient		-0.0043	
Contingency Coefficient		0.0043	

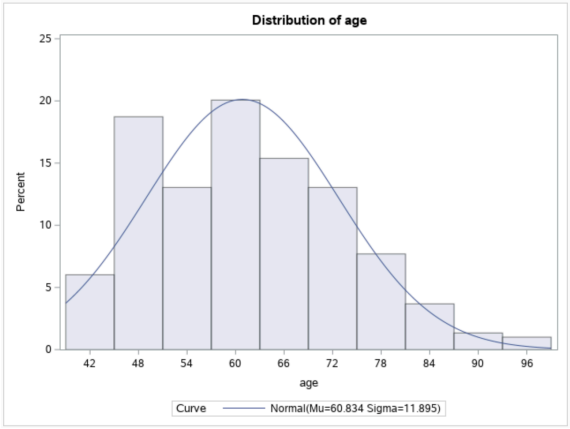
Cramer's V	-0.0043

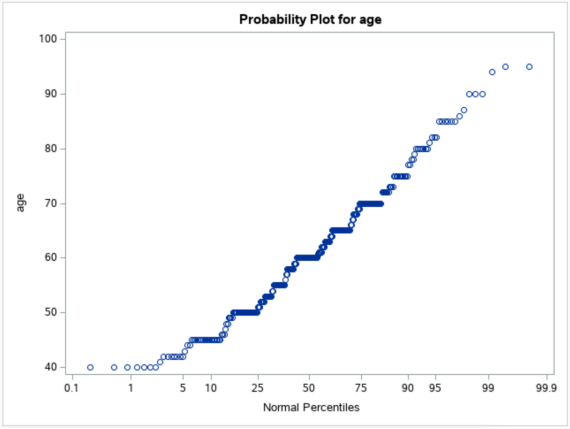
Statistics for Table of Death by smk(Table 3-5)

Statistic	DF	Value	Prob
Chi-Square	1	0.0476	0.8272
Likelihood Ratio Chi-Square	1	0.0478	0.8270
Continuity Adj. Chi-Square	1	0.0073	0.9318
Mantel-Haenszel Chi-Square	1	0.0475	0.8275
Phi Coefficient		0.0126	
Contingency Coefficient		0.0126	
Cramer's V		0.0126	

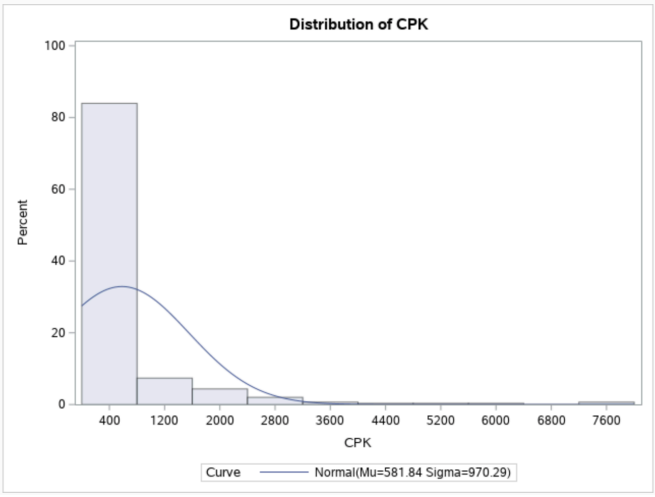
<Table 4-1 ~ 4.7> Goodness-of-fit tests for normality(Continuous variables) + Distribution Plots

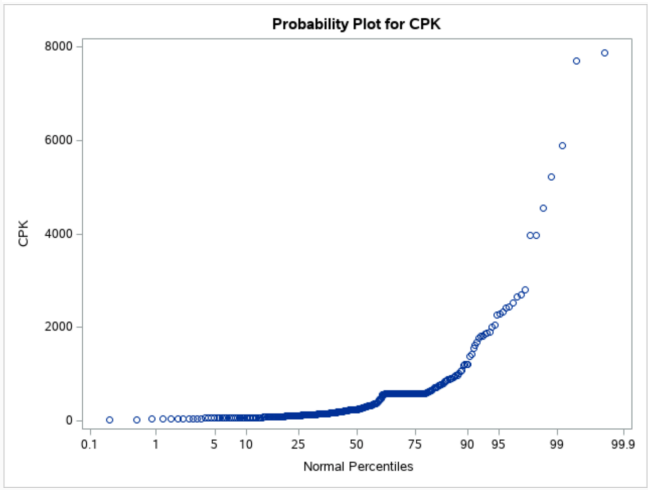
Tests for Normality(age)(Table 4-1)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.97547	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.069751	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.235874	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.642448	**Pr > A-Sq**	<0.0050



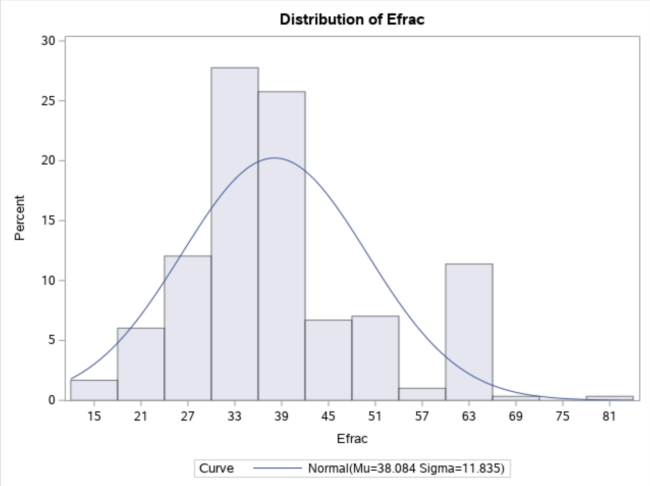


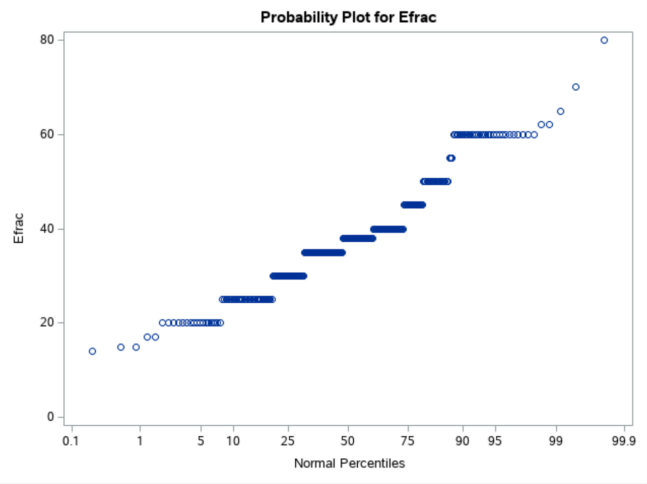
Tests for Normality(CPK)(Table 4-2)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.514263	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.286765	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	8.096309	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	41.90613	**Pr > A-Sq**	<0.0050



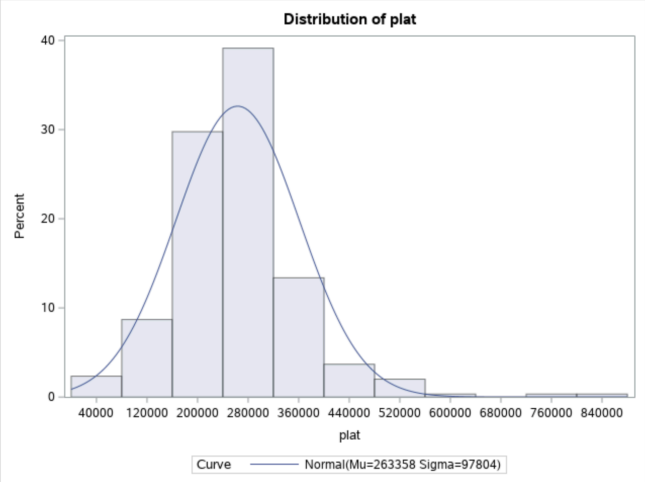


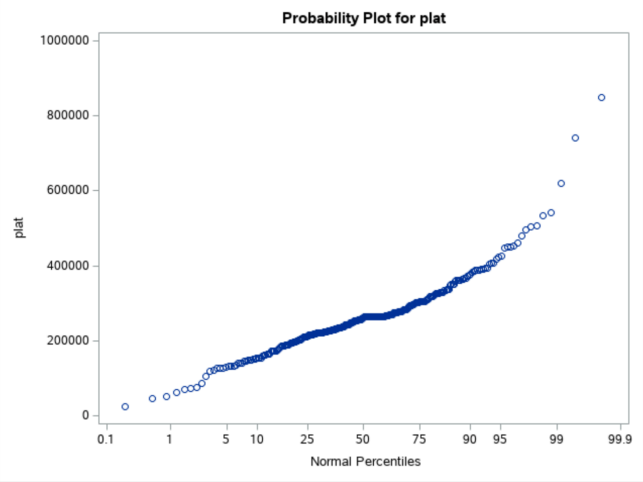
Tests for Normality(Efrac)(Table 4-3)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.947316	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.168123	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.944623	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	5.802017	**Pr > A-Sq**	<0.0050



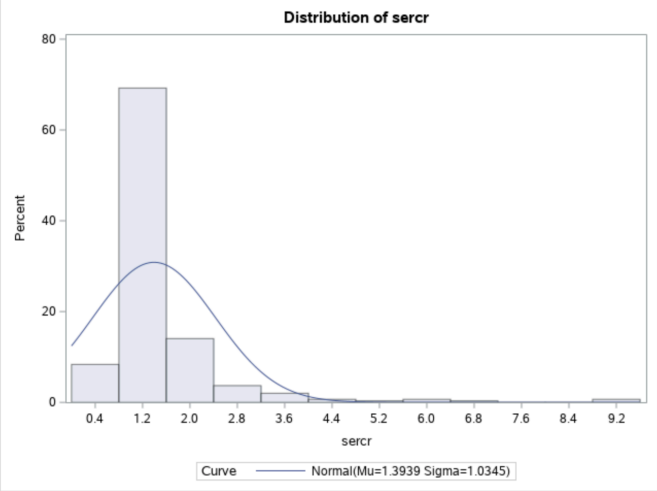


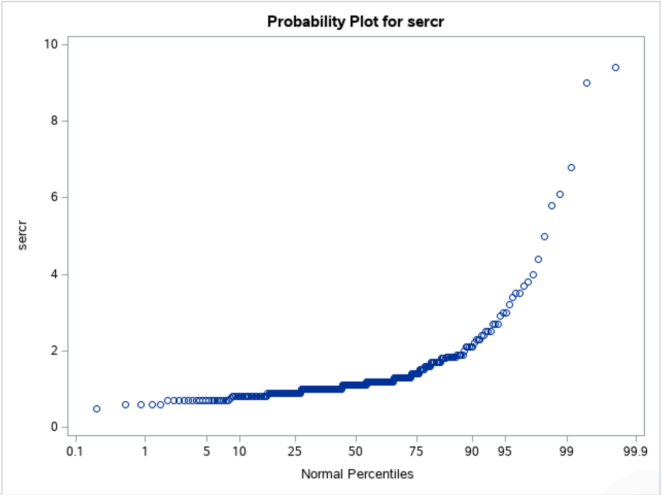
Tests for Normality(plat)(Table 4-4)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.911509	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.116068	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.924776	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	4.989043	**Pr > A-Sq**	<0.0050



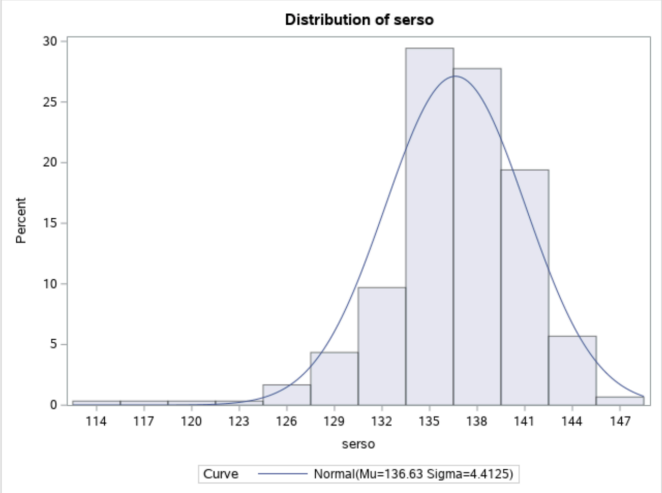


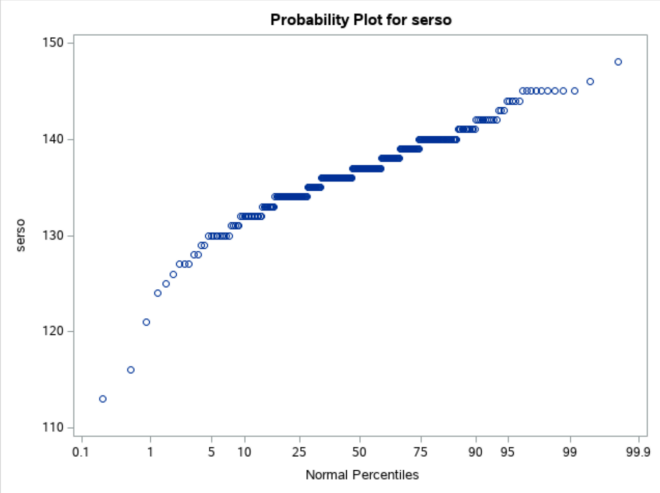
Tests for Normality(sercr)(Table 4-5)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.551466	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.265251	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	6.935724	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	36.45086	**Pr > A-Sq**	<0.0050



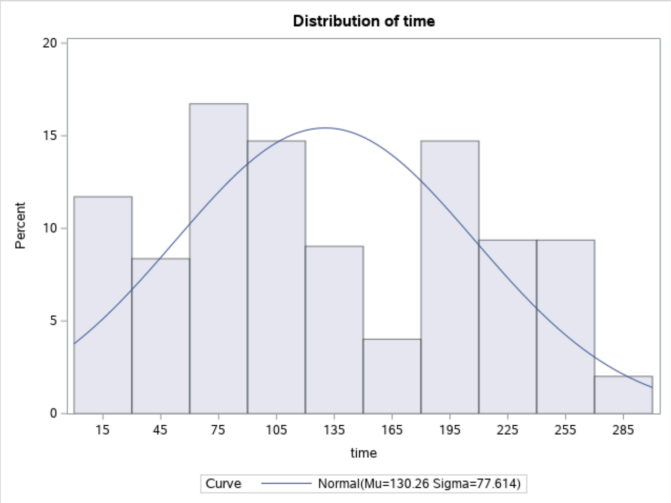


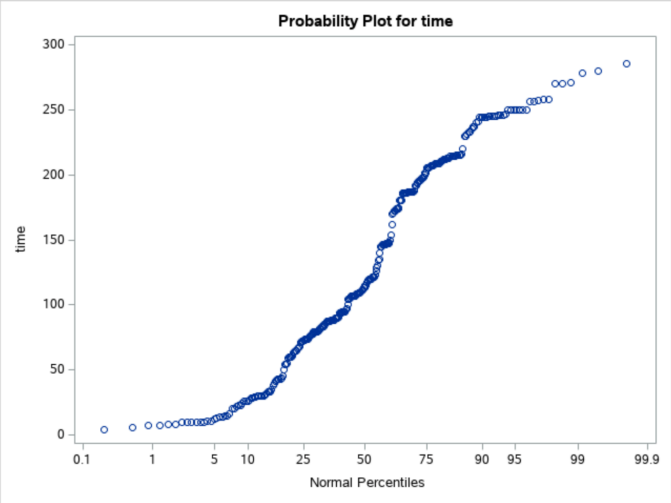
Tests for Normality(serso)(Table 4-6)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.939028	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.11254	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.524928	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	3.093756	**Pr > A-Sq**	<0.0050





Tests for Normality(time)(Table 4-7)				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.946783	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.104807	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.829905	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	4.970228	**Pr > A-Sq**	<0.0050





<Table 5> Mean comparison of continuous variables between death and no death occurrences(Wilcoxon)

The NPAR1WAY Procedure(Table 5-1)

Wilcoxon Scores (Rank Sums) for Variable age Classified by Variable Death					
Death	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
0	203	27827.0	30450.0	696.603384	137.078818
1	96	17023.0	14400.0	696.603384	177.322917

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	**Z**	**Pr > Z**	**Pr > |Z|**	**Pr > Z**	**Pr > |Z|**
17023.00	3.7647	<.0001	0.0002	0.0001	0.0002

The NPAR1WAY Procedure(Table 5-2)

Wilcoxon Scores (Rank Sums) for Variable CPK Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	30166.0	30450.0	696.634859	148.600985
1	96	14684.0	14400.0	696.634859	152.958333

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
14684.00	0.4070	0.3420	0.6840	0.3422	0.6843

The NPAR1WAY Procedure(Table 5-3)

Wilcoxon Scores (Rank Sums) for Variable Efrac Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	33882.50	30450.0	693.137669	166.908867
1	96	10967.50	14400.0	693.137669	114.244792

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
10967.50	-4.9514	<.0001	<.0001	<.0001	<.0001

The NPAR1WAY Procedure(Table 5-4)

Wilcoxon Scores (Rank Sums) for Variable plat Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	31006.50	30450.0	697.778070	152.741379
1	96	13843.50	14400.0	697.778070	144.203125

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
13843.50	-0.7968	0.2128	0.4256	0.2131	0.4262

The NPAR1WAY Procedure(Table 5-5)

Wilcoxon Scores (Rank Sums) for Variable sercr Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	26004.0	30450.0	694.897434	128.098522
1	96	18846.0	14400.0	694.897434	196.312500

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
18846.00	6.3973	<.0001	<.0001	<.0001	<.0001

The NPAR1WAY Procedure(Table 5-6)

Wilcoxon Scores (Rank Sums) for Variable serso Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	32967.50	30450.0	694.991458	162.401478
1	96	11882.50	14400.0	694.991458	123.776042

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
11882.50	-3.6216	0.0001	0.0003	0.0002	0.0003

The NPAR1WAY Procedure(Table 5-7)

Wilcoxon Scores (Rank Sums) for Variable time Classified by Variable Death

Death	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	203	36994.50	30450.0	697.952090	182.238916
1	96	7855.50	14400.0	697.952090	81.828125

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
7855.500	-9.3760	<.0001	<.0001	<.0001	<.0001

<Table 6> The LOGISTIC Procedure for fitting logit model to variables age, Efrac, sercr

The LOGISTIC Procedure

Model Information	
Data Set	WORK.HEARTFAIL
Response Variable	Death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	377.349	313.283
SC	381.049	328.084
-2 Log L	375.349	305.283

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	70.0661	3	<.0001
Score	62.2949	3	<.0001
Wald	49.0745	3	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3531	0.8396	7.8551	0.0051
age	1	0.0517	0.0123	17.6549	<.0001
Efrac	1	-0.0700	0.0142	24.1859	<.0001
sercr	1	0.6659	0.1592	17.5051	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.053	1.028	1.079
Efrac	0.932	0.907	0.959
sercr	1.946	1.425	2.659

<Table 7> Conditional logistic regression by using strata option(stratification being new variable month)

The LOGISTIC Procedure

Conditional Analysis

Model Information	
Data Set	WORK.HEARTFAIL
Response Variable	Death
Number of Response Levels	2
Number of Strata	3
Model	binary logit
Optimization Technique	Newton-Raphson ridge

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
AIC	260.375	218.289
SC	260.375	229.390
-2 Log L	260.375	212.289

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.0859	3	<.0001
Score	42.2041	3	<.0001
Wald	35.3446	3	<.0001

Analysis of Conditional Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
age	1	0.0503	0.0158	10.1507	0.0014
Efrac	1	-0.0806	0.0171	22.3039	<.0001
sercr	1	0.6618	0.1722	14.7681	0.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.052	1.020	1.085
Efrac	0.923	0.892	0.954
sercr	1.938	1.383	2.717

<Figure 6> Oversampling method using PROC SURVEYSELECT

```
data have;
    set heartfail;
run;

proc sort data=have;
    by Death;
run;

proc surveyselect data=have out=want method=urs sampsize=(203 203)  outhits;
    strata Death;
run;
```

<Table 10> Conditional logistic regression by using strata option(stratification being new variable month) with oversampled data

The LOGISTIC Procedure

Conditional Analysis

Model Information	
Data Set	WORK.WANT
Response Variable	Death
Number of Response Levels	2
Number of Strata	3
Model	binary logit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	406
Number of Observations Used	406
Number of Observations Informative	406

Response Profile

Ordered Value	Death	Total Frequency
1	1	203
2	0	203

Probability modeled is Death=1.

Strata Summary

Response Pattern	Death 1	Death 0	Number of Strata	Frequency
1	50	3	1	53
2	57	2	1	59
3	96	198	1	294

Newton-Raphson Ridge Optimization Without Parameter Scaling

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	400.440	301.511
SC	400.440	313.530
-2 Log L	400.440	295.511

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	104.9294	3	<.0001
Score	86.3294	3	<.0001
Wald	65.7111	3	<.0001

Analysis of Conditional Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
age	1	0.0632	0.0134	22.3095	<.0001
Efrac	1	-0.0961	0.0155	38.3765	<.0001
sercr	1	0.9550	0.1855	26.5097	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.065	1.038	1.093
Efrac	0.908	0.881	0.936
sercr	2.599	1.807	3.738