南方科技大学 Southern University of Science and Technology | 统计与数据科学系 Department of Statistics and Data Science

Survival Analysis of Patients with Arrhythmia

Author: Feng Zhen 11711135

Liu Runqi 11711331

Wu Siqi 11711114

Course: MA405 Survival Analysis

Instructor: Cong, Xu

Date: Dec. 21, 2020

Abstract

Cardiac arrhythmia refers to a medical condition in which heart beats irregularly. One possible treatment for patients who suffer from arrhythmia is to install a device called Implantable Cardioverter Defibrillator (ICD). This report aims to identify high-risk patients most in need of an ICD based on the arrhythmia dataset with 27 predictors. Based on exploratory analysis, stepwise variable selection and other previous studies, a few models including Cox regression model and parametric AFT model are built. After model optimization and verification, we obtain significant variables that contribute to high-risk in arrhythmia and demand for implanting ICD.

Keywords: arrhythmia, ICD, survival analysis, Cox regression model, AFT model

1. Introduction

Arrhythmias are abnormal beats. The term "arrhythmia" refers to any change from the normal sequence of electrical impulses, causing abnormal heart rhythms. Arrhythmias may be completely harmless or life-threatening. In the US, Arrhythmia contributes to approximately 200,000-300,000 sudden deaths per year – a higher incidence than stroke, lung cancer or breast cancer.[1] An implantable cardioverter-defibrillator (ICD) is a battery-powered device placed under the skin that detects and stops abnormal heartbeats. The device continuously monitors patient's heartbeat and delivers electrical pulses to restore a normal heart rhythm when necessary.

The data in this study are from a large, multi-hospital study of patients with ICDs. Overall survival (time until death) and the time until the ICD delivered its first shock are two primary incomes. A variety of potentially useful predictors, including demographic, laboratory values, medical history, and electrocardiogram (ECG)-derived variables, are available. The goals of the study are to detect the predictors that significantly influence the risk of abnormal heartbeats and construct efficient models for identifying high-risk patients most in need of an ICD.

2. Data processing and analysis

The data in the study contains 946 observations and 27 predictors such as age, sex, blood pressure, and heart rate. 16 predictors are categorical and the rest are numeric. Four response variables are indicators for shock and death (0 for censor and 1 for happen), and the time when shock and death happen or censor (in months).

2.1 Data processing

Import the data into R and no missing value is detected. Then, for the convenience of analysis, we assign value to the categorical variables. For instance, Sex=0 for female and Sex=1 for male, Diagnosis=0 for Idiopathic and Diagnosis=1 for Ischemic, 1 to 6 representing six different races, 0 for No and 1 for Yes, etc.

There is an abnormal observation who died at 41.6 months but his shock time censored at 51.1 months. Given this situation is impossible to happen, we remove the abnormal observation.

2.2 Demographic statistics

Most observations are older than 50 years, and the youngest age, the oldest age, and the median are 20, 87, and 64 years, respectively. *Figure 1* demonstrates the distribution of age. *Table 1* summarizes the race and gender of the patients: Most observations are white, followed by African American and there are much more males than females. New York Heart Association classification (NYHA) is the most common measure of heart failure severity, where class I and class II are mild cardiac disease, class III is moderate and class IV is severe. Figure 2 shows the distribution of NYHA: most patients with ICD are in class 2 (549) and class 3 (258), and only 5 patients are in class 4.

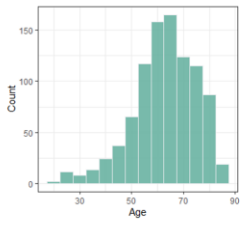

Figure 1

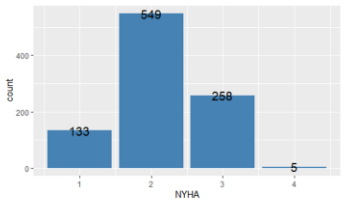

Figure 2

	African American	American Indian	Asian	White	Other	No answer	Total
Female	50	1	0	138	0	2	191
Male	111	2	5	621	2	13	754
Total	161	3	5	759	2	15	945

Table 1

2.3 Classification

In this section, we aim to classify the patients by whether they need ICD based on the information from shock and death. Specifically, if a patient died within one day after shock, we consider ICD is useless and classify the patient to Group 1; if a patient's shock censors at short time (in this study, we take tDeath-tShock≤0.01month) before he/she dies, we consider shocks have never happen to him/her during the study so the individual does not need ICD and is classified to Group 2; if shock once happened to a patient, and the patient lives longer than one day after shock, then we consider ICD is an effective treatment for him/her and classify him/her to Group 3. Hence, both Group1 and Group 2 do not need ICD but Group 3 need ICD. Based on above standard, we are able to classify 195 observations, 3 in Group 1, 5 in Group 2, and 187 in Group 3. Because there are too few observations in Group 1 comparing to Group 3, ICD can save one's life through shock, when abnormal heart rhythms happen. Therefore, almost all patients with high risk of shock need ICDs. Thus, in the study, we need to identify the patients with high risk of Arrhythmias (Shock is an indicator for Arrhythmias).

3. Exploratory analysis

3.1 Correlation

Figure 3 illustrates the correlation between numeric variables, deeper color indicating the stronger correlation between two variables. There are three pairs of variables show relatively strong correlation: cor(SysBP,DiaBP)=0.56, cor(BUN,Creatinine)=0.58, cor(QRS,QTc)=0.54. It is trivial that SysBP(Systolic blood pressure) and DiaBP (Diastolic blood pressure) are closely related. However, variables with high correlation coefficient do not necessarily have strong relationship. To interpret the possible relationship between variables, further

analysis and investigation are needed.

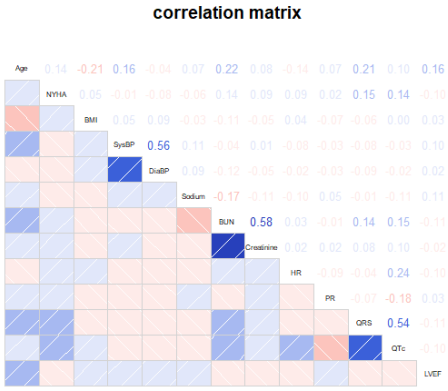

Figure 3

3.2 Comparison

To identify the patients who need ICD most, an effective approach is to compare the arrhythmias risk of patients with different features. In this section, we compare the Kaplan-Meier estimate curve for shock of patients in different groups. Grouping is based on the observation's value of a certain variable. For categorical variables, grouping standard is trivial—patients with the same value are in the same group. For numeric variables, we apply different grouping method. For variable "Age", we divide observations into four groups based on quartile. For the variables that have normal range, including BMI (Body mass index), blood pressure and heart rate, we divide observations into three groups: Low, Normal, and High. There are three particular variables, QRS (Width of the QRS complex, from ECG), QTc (Corrected QT interval, from ECG) and LVEF (Left ventricular ejection fraction) should be noted. The normal range of QRS is 70 to 100. However, the QRS of over three quarters of sample is greater than 100. As a result, we divide the sample based on whether QRS value exceeds 100. Similarly, the normal range of QTc is (400, 440), and we divide the sample based on whether QTc value exceeds 440. The normal range of LVEF is 55 to 70. However, the greatest LVEF in the sample is 38, and the median is 20. Therefore, the sample are divided into two groups by comparing values with median. A possible explanation for the phenomenon is that these patients were implanted with ICD because they had the risk of arrhythmia, and patients with high QRS, high QTc and low LVEF tend to have cardiac failure [3][4][5].

Figure 4 (a)-(c) show the estimate survival function for shock of people in different age groups, sex groups and race groups. The patients younger than 56 years

old have the greatest risk of shock while the patients between 63 to 72 years old have the lowest risk. Male patients have significantly higher risk of shock than females. Also, since the sample sizes of American Indian and Asian are too small, we cannot reach conclusions about them from curves. From *Figure 4 (c)* African American patients have higher risk of arrhythmias than white patients. *Figure 4 (d)* suggests patients with ischemic cardiomyopathy have slightly higher risk of shock. *Figure 4 (e)* suggests that patients in severe group have the highest risk while the risks in other groups do not differ much. *(f)* and *(g)* reveal the same fact: patients with high blood pressure suffer higher risk of arrhythmias, so the two predictors are related. *(h)-(k)* have the same pattern: two curves do not intersect. Given the fact that these variables are indicators of taking medication and most patients take at least two of them simultaneously, these variables are likely correlated. Patients with high BUN (blood urea nitrogen) do not show higher risk than patients in normal group. Patients with high Creatinine (serum creatine) show slightly higher risk than normal patients. From *Figure 4 (l)* and *(m)*, we cannot see the relation between BUN and Creatinine. Similarly, *Figure 4 (n)* and *(o)* do not reveal obvious relation between QRS and QTc. *Figure 4 (p)* shows that in the short term, patients with low LVEF have higher risk of arrhythmias while in the long term (t>70 months), patients which higher but not normal LVEF level have higher risk of arrhythmias.

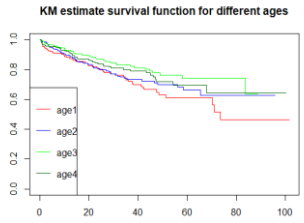

Figure 4 (a)

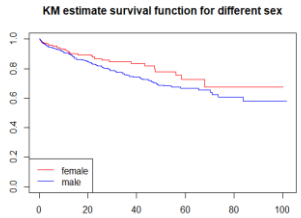

Figure 4 (b)

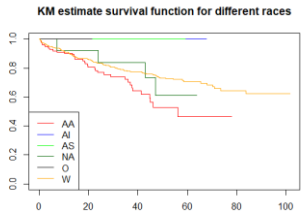

Figure 4 (c)

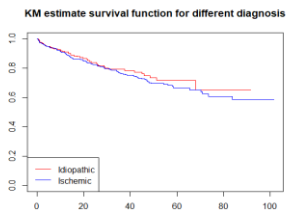

Figure 4 (d)

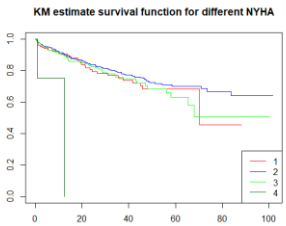

Figure 4 (e)

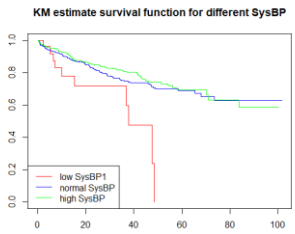

Figure 4 (f)

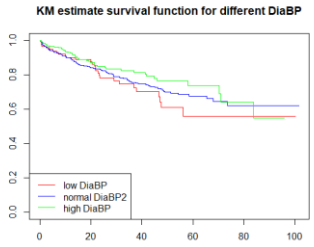

Figure4(g)

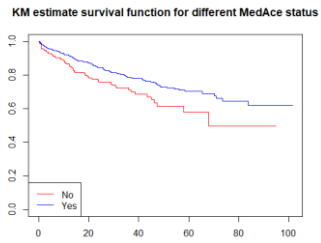

Figure 4 (h)

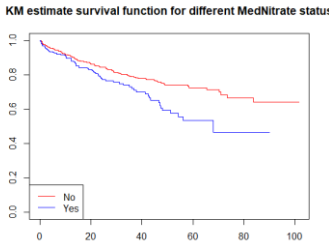

Figure 4 (i)

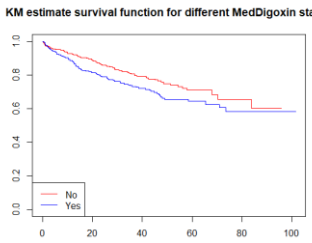

Figure 4 (j)

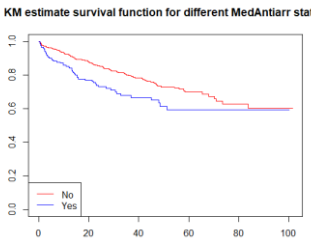

Figure 4 (k)

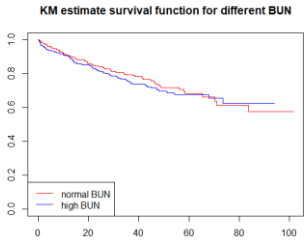

Figure 4 (l)

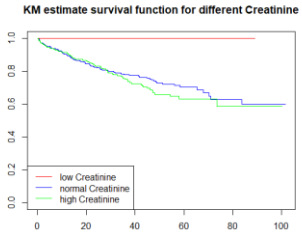

Figure 4 (m)

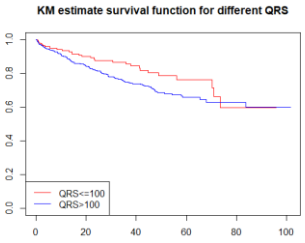

Figure 4 (n)

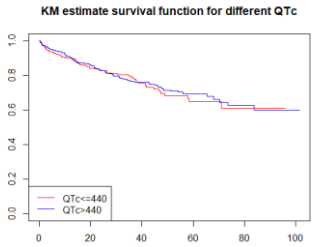

Figure 4 (o)

4. Statistical modeling

After classifying all patients into three groups in section 2.3, we found that there are very few patients whose ICD was useless. Therefore, in the following process, we consider all individuals in the dataset under assumption that their ICDs were working, even though some of them might not need one. In this case, our goal is to identify what kinds of patients have a high risk of Arrhythmias, in another word, need an ICD essentially. Two important indicators for risk of Arrhythmias are whether and when ICD deliver electric shock ("dShock" and "tShock").

4.1 Overall estimated survival and hazard functions

First, we took a look at the general survival probability and hazard rate of both shock and death.

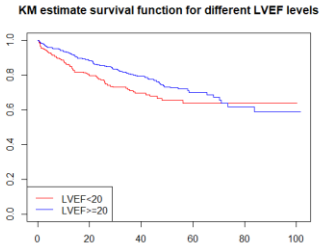

Figure 4 (p)

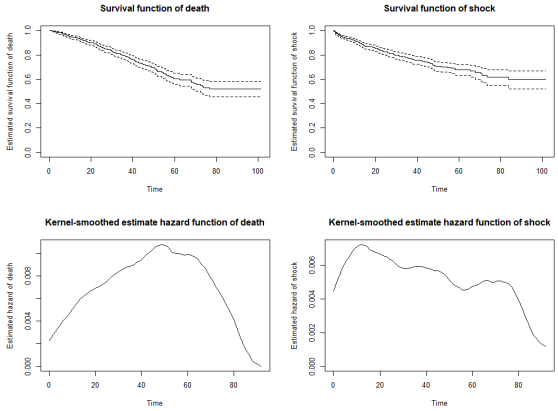

Figure 5

From the figure above, we conclude that hazard of death is not statistically related to hazard of shock. Thus, death will not be considered as a component when fitting the model. We decided to fit both Cox Proportional model and AFT model with dependent variables as "dShock" and "tShock", then using statistic methods to determine which one is better.

4.2 Variable selection and Cox Proportional model

Our variable selection procedure is mainly based on exploratory analysis result, correlation matrix and scientific facts. First, fit a full model with all 27 variables. Second, we subjectively select a few variables, based on the result in exploratory analysis and correlation matrix in *Figure 3*. In exploratory analysis, variables that we consider as significant are: Age, Sex, Race, Diagnosis, NYHA, HxHTN, HxChol, BMI, SysBP, DiaBP, MedAce, MedNitrate, MedDigoxin, MedAntiarr, BUN, Creatinine, PR, QRS, LVEF. Among these variables, we also detected three pairs highly correlated variables from the result of correlation matrix, which are (SysBP, DiaBP), (BUN, Creatinine), (QRS, LVEF). The final subjective model includes only one component from each pair, SysBP, BUN and LVEF. Then, apply stepwise selection procedure on full model and get a model with smallest AIC value.

```
Call:
coxph(formula = Surv(tShock, dShock) ~ Age + Sex + Race + Diagnosis +
    NYHA + HxMI + HxDiabetes + HxHTN + HxChol + BMI + SysBP +
    DiaBP + MedAce + MedBeta + MedNitrate + MedDiuretic + MedAldactone +
    MedDigoxin + MedAntiarr + Sodium + BUN + Creatinine + HR +
    PR + QRS + QTc + LVEF, data = data)

  n= 945, number of events= 198

                coef  exp(coef)   se(coef)      z Pr(>|z|)
Age         -0.0181474  0.9820162  0.0070628 -2.569 0.010187 *
Sex          0.3105185  1.3641323  0.2170089  1.431 0.152458
Race        -0.0754727  0.9273051  0.0431461 -1.749 0.080250 .
Diagnosis    0.1837822  1.2017541  0.2378149  0.773 0.439644
NYHA        -0.0439786  0.9569744  0.1175735 -0.374 0.708366
HxMI         0.0217460  1.0219841  0.1920377  0.113 0.909842
HxDiabetes  -0.1442349  0.8656844  0.1722662 -0.837 0.402436
HxHTN       -0.1200296  0.8868942  0.1640789 -0.732 0.464452
HxChol       0.0846733  1.0883615  0.1689495  0.501 0.616248
BMI          0.0099564  1.0100061  0.0137519  0.724 0.469065
SysBP       -0.0037898  0.9962174  0.0052340 -0.724 0.469020
DiaBP       -0.0054158  0.9945989  0.0075331 -0.719 0.472182
MedAce      -0.3649109  0.6942585  0.1721491 -2.120 0.034028 *
MedBeta      0.1641871  1.1784347  0.2158319  0.761 0.446826
MedNitrate   0.3669062  1.4432625  0.1662881  2.206 0.027353 *
MedDiuretic  0.2697298  1.3096106  0.1768109  1.526 0.127128
MedAldactone 0.1608789  1.1745427  0.1680093  0.958 0.338285
MedDigoxin   0.3402897  1.4053546  0.1534984  2.217 0.026630 *
MedAntiarr   0.4627084  1.5883701  0.1740570  2.658 0.007852 **
Sodium      -0.0124304  0.9876465  0.0222266 -0.559 0.575984
BUN          0.0128062  1.0128886  0.0062947  2.034 0.041906 *
Creatinine   0.0060379  1.0060562  0.0993792  0.061 0.951553
HR          -0.0008825  0.9991179  0.0051188 -0.172 0.863122
PR           0.0038607  1.0038682  0.0010881  3.548 0.000388 ***
QRS          0.0015733  1.0015745  0.0025608  0.614 0.538969
QTc         -0.0005736  0.9994266  0.0018491 -0.310 0.756419
LVEF        -0.0214190  0.9788087  0.0126296 -1.696 0.089897 .
---
```

Figure 6 (Full model)

```
Call:
coxph(formula = Surv(tShock, dShock) ~ Age + Sex + Race + Diagnosis +
    NYHA + HxHTN + HxChol + BMI + DiaBP + MedAce + MedNitrate +
    MedDigoxin + MedAntiarr + BUN + PR + LVEF, data = data)

  n= 945, number of events= 198

                coef exp(coef)  se(coef)      z Pr(>|z|)
Age         -0.019005  0.981174  0.006565 -2.895 0.003790 **
Sex          0.336580  1.400151  0.213300  1.578 0.114573
Race        -0.079227  0.923830  0.041603 -1.904 0.056861 .
Diagnosis    0.156054  1.168890  0.197030  0.792 0.428342
NYHA        -0.042568  0.958326  0.117051 -0.364 0.716106
HxHTN       -0.134035  0.874559  0.161302 -0.831 0.405997
HxChol       0.071613  1.074239  0.166916  0.429 0.667898
BMI          0.010070  1.010121  0.012926  0.779 0.435934
DiaBP       -0.010316  0.989737  0.006172 -1.671 0.094635 .
MedAce      -0.335731  0.714815  0.168353 -1.994 0.046129 *
MedNitrate   0.351861  1.421711  0.159348  2.208 0.027235 *
MedDigoxin   0.366225  1.442280  0.148211  2.471 0.013475 *
MedAntiarr   0.475469  1.608768  0.165272  2.877 0.004016 **
BUN          0.012508  1.012586  0.004747  2.635 0.008414 **
PR           0.003925  1.003932  0.001061  3.698 0.000217 ***
LVEF        -0.024119  0.976170  0.012199 -1.977 0.048023 *
---
```

Figure 7 (Subjective model)

```
Call:
coxph(formula = Surv(tShock, dShock) ~ Age + Sex + Race + DiaBP +
    MedAce + MedNitrate + MedDiuretic + MedDigoxin + MedAntiarr +
    BUN + PR + LVEF, data = data)

  n= 945, number of events= 198

                coef exp(coef)  se(coef)      z Pr(>|z|)
Age         -0.019677  0.980515  0.005998 -3.281 0.001036 **
Sex          0.415502  1.515131  0.201950  2.057 0.039643 *
Race        -0.060835  0.940979  0.039702 -1.532 0.125456
DiaBP       -0.009845  0.990204  0.006069 -1.622 0.104795
MedAce      -0.383200  0.681677  0.167618 -2.286 0.022246 *
MedNitrate   0.344103  1.410724  0.155530  2.212 0.026935 *
MedDiuretic  0.306818  1.359094  0.171704  1.787 0.073953 .
MedDigoxin   0.337156  1.400957  0.148747  2.267 0.023412 *
MedAntiarr   0.460649  1.585103  0.164188  2.806 0.005022 **
BUN          0.011130  1.011193  0.004767  2.335 0.019553 *
PR           0.003854  1.003861  0.001047  3.681 0.000232 ***
LVEF        -0.023260  0.977009  0.012175 -1.911 0.056067 .
---
```

Figure 8 (Stepwise model)

Comparing subjective model with stepwise model, majority variables are the same except Diagnosis, NYHA, HxHTN, HxChol and BMI. A study in 2015 indicated that Arrhythmias frequently happen on both ischemic cardiomyopathy and non–ischemic cardiomyopathy patients, so Diagnosis should not include in the final model [6]. In a research in 2010, doctors concluded that, "*In chronic ischemic patients with an ICD for primary prevention, the presence of diabetes, renal dysfunction, higher NYHA class, and impaired peri-infarct zone function were predictors of all-cause mortality*" [7]. Thus, NYHA should be included in the final model. For HxHTM and HxChol, a study in 2002 introduced that Arrhythmias are common problems in hypertensive patients [8], while the relation between high cholesterol and Arrhythmias can be controversial [9][10]. Besides, "*Atrial fibrillation (AF) is commonly associated with overweight and obesity*", said by European researchers in 2016 [11]. Therefore, BMI should be included in the final model. Finally, we have the final cox model.

```
Call:
coxph(formula = Surv(tShock, dShock) ~ Age + Sex + Race + NYHA +
    HxHTN + BMI + DiaBP + MedAce + MedNitrate + MedDigoxin +
    MedAntiarr + BUN + PR + LVEF, data = data)

  n= 945, number of events= 198

                coef exp(coef)  se(coef)      z Pr(>|z|)
Age         -0.017111  0.983035  0.006248 -2.739 0.006168 **
Sex          0.400545  1.492638  0.202331  1.980 0.047742 *
Race        -0.071164  0.931309  0.040776 -1.745 0.080941 .
NYHA        -0.040401  0.960405  0.116927 -0.346 0.729704
HxHTN       -0.119078  0.887739  0.159316 -0.747 0.454803
BMI          0.010641  1.010698  0.012815  0.830 0.406348
DiaBP       -0.010682  0.989375  0.006205 -1.722 0.085136 .
MedAce      -0.352652  0.702822  0.167576 -2.104 0.035341 *
MedNitrate   0.380611  1.463179  0.156532  2.432 0.015036 *
MedDigoxin   0.369260  1.446663  0.148341  2.489 0.012801 *
MedAntiarr   0.470385  1.600611  0.165094  2.849 0.004383 **
BUN          0.012083  1.012156  0.004745  2.546 0.010885 *
PR           0.003865  1.003872  0.001063  3.636 0.000277 ***
LVEF        -0.023362  0.976909  0.012171 -1.919 0.054930 .
---
```

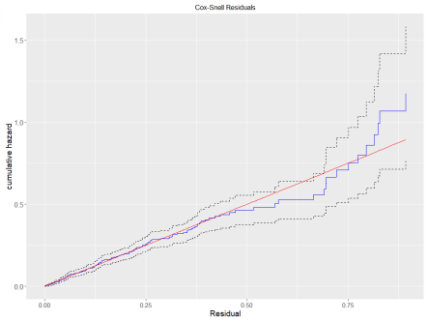

Figure 9 (Final Cox model) *Figure 10 (Cox-Snell residual plot of final Cox model)*

4.3 AFT model

Now we want to fit an AFT model in case that the proportional hazard assumption does not hold. From hazard plot of shock in *Figure 5*, it is not a unimodal function, so log-logistic distribution may not be appropriate. Anyway, we will fit AFT model with all three common distributions, Weibull, Log-logistic and Log-normal.

```
Call:
survreg(formula = Surv(tShock, dShock) ~ Age + Sex + Race + NYHA +
    HxHTN + BMI + DiaBP + MedAce + MedNitrate + MedDigoxin +
    MedAntiarr + BUN + PR + LVEF, data = data, dist = "weibull")
              Value Std. Error     z       p
(Intercept)  3.95152    0.94407  4.19 2.8e-05
Age          0.02002    0.00745  2.69 0.00720
Sex         -0.48686    0.24185 -2.01 0.04410
Race         0.08332    0.04826  1.73 0.08427
NYHA         0.04788    0.13849  0.35 0.72952
HxHTN        0.14546    0.18919  0.77 0.44196
BMI         -0.01314    0.01519 -0.87 0.38694
DiaBP        0.01268    0.00740  1.71 0.08687
MedAce       0.42202    0.20024  2.11 0.03507
MedNitrate  -0.44827    0.18561 -2.42 0.01573
MedDigoxin  -0.43745    0.17766 -2.46 0.01380
MedAntiarr  -0.55795    0.19713 -2.83 0.00465
BUN         -0.01423    0.00561 -2.54 0.01118
PR          -0.00461    0.00126 -3.64 0.00027
LVEF         0.02750    0.01447  1.90 0.05733
Log(scale)   0.17028    0.06136  2.78 0.00552
```

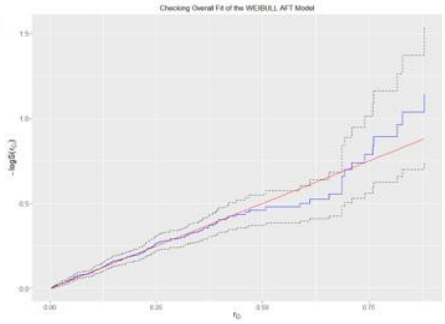

Figure 11 (Weibull AFT model)

```
Call:
survreg(formula = Surv(tShock, dShock) ~ Age + Sex + Race + NYHA +
    HxHTN + BMI + DiaBP + MedAce + MedNitrate + MedDigoxin +
    MedAntiarr + BUN + PR + LVEF, data = data, dist = "loglogistic")
              Value Std. Error     z      p
(Intercept)  3.50116    1.03242  3.39 0.0007
Age          0.02031    0.00809  2.51 0.0121
Sex         -0.51811    0.25031 -2.07 0.0385
Race         0.07833    0.05129  1.53 0.1267
NYHA         0.04444    0.14793  0.30 0.7639
HxHTN        0.16509    0.20281  0.81 0.4156
BMI         -0.01435    0.01638 -0.88 0.3813
DiaBP        0.01309    0.00787  1.66 0.0963
MedAce       0.50548    0.21663  2.33 0.0196
MedNitrate  -0.39557    0.20155 -1.96 0.0497
MedDigoxin  -0.48565    0.18773 -2.59 0.0097
MedAntiarr  -0.62015    0.21200 -2.93 0.0034
BUN         -0.01497    0.00661 -2.26 0.0236
PR          -0.00429    0.00151 -2.84 0.0046
LVEF         0.03167    0.01541  2.06 0.0398
Log(scale)   0.07721    0.06115  1.26 0.2067
```

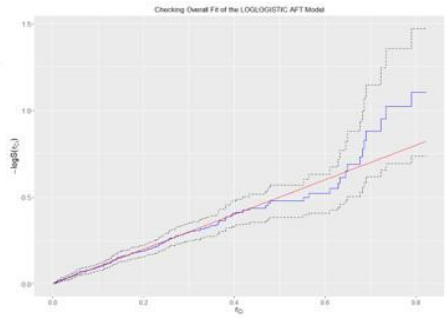

Figure 12 (Log-logistic AFT model)

```
Call:
survreg(formula = Surv(tShock, dShock) ~ Age + Sex + Race + NYHA +
    HxHTN + BMI + DiaBP + MedAce + MedNitrate + MedDigoxin +
    MedAntiarr + BUN + PR + LVEF, data = data, dist = "lognormal")
              Value Std. Error     z       p
(Intercept)  3.62338    1.16006  3.12  0.0018
Age          0.02283    0.00909  2.51  0.0120
Sex         -0.53796    0.26786 -2.01  0.0446
Race         0.07801    0.05640  1.38  0.1666
NYHA         0.02899    0.15909  0.18  0.8554
HxHTN        0.25762    0.22250  1.16  0.2469
BMI         -0.01396    0.01800 -0.78  0.4380
DiaBP        0.01511    0.00868  1.74  0.0818
MedAce       0.55130    0.23816  2.31  0.0206
MedNitrate  -0.43103    0.22157 -1.95  0.0517
MedDigoxin  -0.47165    0.20322 -2.32  0.0203
MedAntiarr  -0.64812    0.23355 -2.78  0.0055
BUN         -0.01862    0.00741 -2.51  0.0120
PR          -0.00508    0.00183 -2.78  0.0055
LVEF         0.03191    0.01674  1.91  0.0565
Log(scale)   0.75809    0.05505 13.77 <2e-16
```

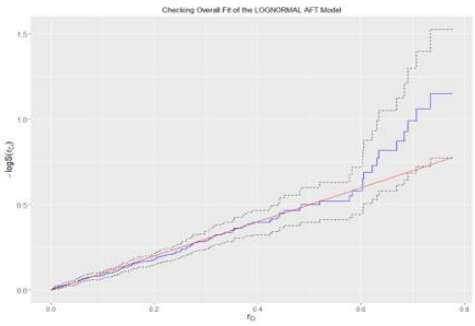

Figure 13 (Log-normal AFT model)

Overall, we will use Weibull AFT model as the final AFT model, then compare the final AFT model with the final Cox model.

5. Model Verification

5.1 Model Diagnosis for the Cox PH Model

To check the proportional hazards assumption, we draw scaled Schoenfeld residuals plots for each coefficient in the model. A horizontal line in the plot suggests the coefficient of X_j is constant, which means the proportional hazards assumption is satisfied. From figure 14, we find out that variables MedAntiarr and MedDigoxin show strong evidence of nonproportionality. From figure 15, there exist outliers when observations fit the final Cox model.

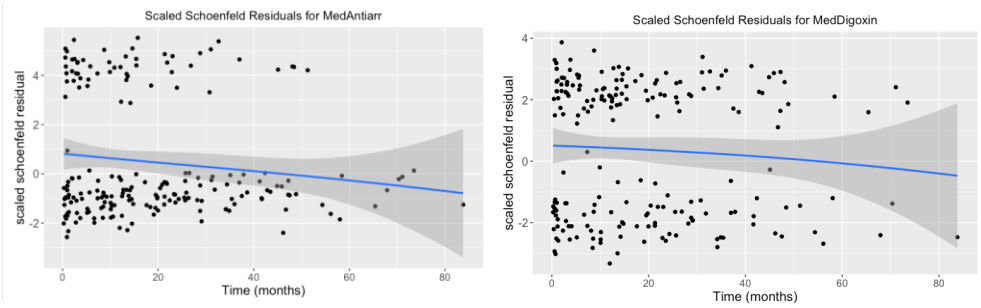

Figure 14 (Scaled Schoenfeld Residuals plot)

To make sure nonproportionality matters, we first check whether it is due to small numbers of outliers by dropping the outliers whose deviance residuals are beyond 3 or below -3.

Index	587	673	793	844	919
Residuals (D)	3.0183	3.1855	3.1176	3.2241	3.4367

Table 2 (Deviance Residuals for outliers)

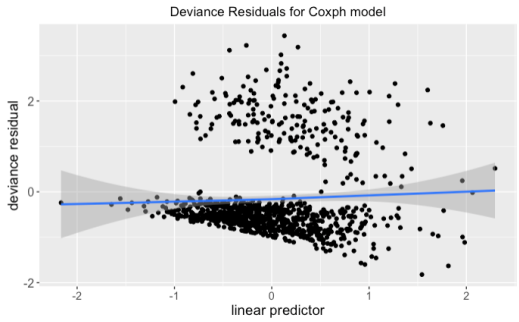

Figure 15 (Deviance Residuals plot)

However, after dropping the outliers, the scaled Schoenfeld residual plots still show that there exist nonproportional effects for these two variables. Thus we consider changing these covariates into stratification factors strata(MedDigoxin) and strata(MedAntiarr) to this model. From figure 17 we find that the proportional hazards assumption holds. We denote this model as Cox model 1 as figure 16.

```
Call:
coxph(formula = Surv(tShock, dShock) ~ Age + Sex + Race + Diagnosis +
    NYHA + HxHTN + DiaBP + MedAce + MedNitrate + strata(MedDigoxin) +
    strata(MedAntiarr) + BUN + PR + LVEF, data = data_2)

               coef exp(coef)  se(coef)      z        p
Age       -0.018736  0.981439  0.006438 -2.910 0.003611
Sex1       0.348980  1.417620  0.213493  1.635 0.102129
Race      -0.081512  0.921722  0.041472 -1.965 0.049358
Diagnosis1  0.169083  1.184218  0.190815  0.886 0.375558
NYHA      -0.021657  0.978576  0.115869 -0.187 0.851734
HxHTN     -0.110102  0.895743  0.158009 -0.697 0.485923
DiaBP     -0.009889  0.990160  0.006196 -1.596 0.110482
MedAce    -0.357357  0.699523  0.168422 -2.122 0.033855
MedNitrate  0.328452  1.388816  0.160310  2.049 0.040477
BUN        0.011430  1.011495  0.004762  2.400 0.016390
PR         0.003594  1.003600  0.001036  3.469 0.000522
LVEF      -0.022302  0.977945  0.012290 -1.815 0.069588

Likelihood ratio test=52.17  on 12 df, p=5.784e-07
n= 940, number of events= 198
```

Figure 16 (Cox model 1)

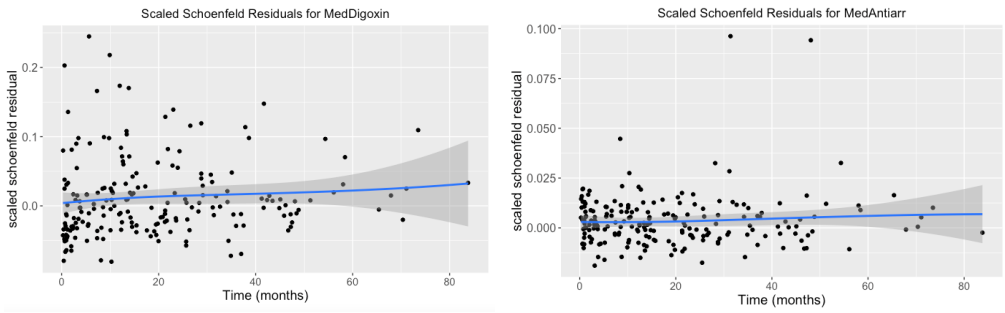

Figure 17 (Scaled Schoenfeld Residuals plots)

In order to investigate whether the correct functional forms for the variates have been used, martingale residuals are calculated and plotted against the values of continuous variables. From figure 18, the explanatory plot of BUN suggests a piecewise linear model with a changepoint at 80. We denote this model as Cox model 2.

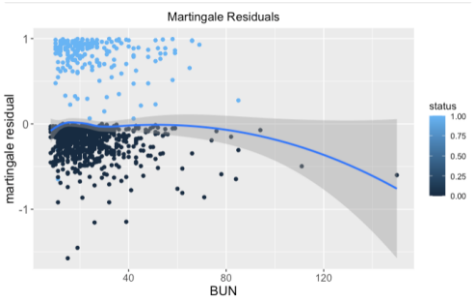

Figure 18 (Martingale Residuals plot)

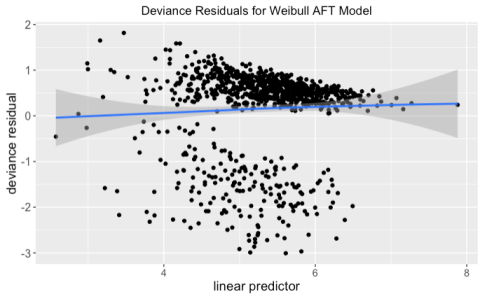

Figure 19 (final AFT model)

5.2 Model Diagnosis for AFT Model

The Weibull AFT model obtained from the statistical modeling process is denoted as final AFT model. We calculate the deviance residuals and find out observations that correspond to relatively deviance residuals (beyond ± 3). Those observations are not well fitted by final AFT model thus they serve as outliers.

Index	844	919
Residuals (D)	-3.0036	-3.0003

Table 3 (Deviance residuals for AFT model 1)

Since the deviance residuals plotted against the linear predictors show no systematic pattern and most observations are in normal range, the final AFT model is overall a good fit.

5.3 Model comparison

Stratified by MedDigoxin and MedAntiarr and adding a piecewise linear term for BUN, Cox model 2 contains 17 variables, while AFT model 1 contains 18 variables. We find out that the values of the coefficients are extremely similar within each pair, since either the Cox regression model or the Weibull distribution AFT model is reasonable by verifications above. But if we compare the coefficients between the two models and calculate the log-likelihoods, the log-likelihood of Cox model 2 is larger. Since the model with higher log-likelihood owns better parametric fitting function, we should choose Cox model 2 as the final model.

	coef_cox
Age	-0.017953541
SexM	0.390655218
RaceAmerican Indian	-15.990914468
RaceAsian	-15.736615728
RaceNo answer	-0.426650129
RaceOther	-14.805082405
RaceWhite	-0.503188997
BMI	0.014474408
NYHA	-0.062020141
HxHTNYes	-0.133774327
DiaBP	-0.009887718
MedAceYes	-0.359039016
MedNitrateYes	0.320269613
BUN	-0.175184799
func1(BUN)	0.193591338
PR	0.003635482
LVEF	-0.026173216

	coef_weibull
Age	-0.016690608
SexM	0.405487848
RaceAmerican Indian	-16.645690231
RaceAsian	-16.368214894
RaceNo answer	-0.397503591
RaceOther	-15.363495386
RaceWhite	-0.397509980
BMI	0.012285132
NYHA	-0.039196122
HxHTNYes	-0.143273086
DiaBP	-0.010378076
MedAceYes	-0.364821218
MedNitrateYes	0.377006696
MedDigoxinYes	0.371650581
MedAntiarrYes	0.475713339
BUN	0.011759481
PR	0.003826375
LVEF	-0.023985629

Figure 20 (comparison of two model's coefficients)

Model	Log-likelihood
Cox model 2	-953.9214
AFT model 1	-1172.991

Table 4 (comparison of log-likelihood)

5.4 Results and interpretation

Finally, we obtain the results of characteristics that could identify high-risk patients most in need of an ICD. In this dataset, patients in dire need of ICD are more likely to age in the first two groups based on quartile and have higher BMI than the average level. White males are in higher risk than other crowds. The majority of patients who need ICD are classified as rank I and II from the New York Heart Association classification. Patients without a history of hypertension are more suitable than those who have the medical history to implement the ICD. Those who have lower diastolic blood pressure and lower left ventricular ejection fraction are also more befitting to use ICD to prevent death. As for medication taking, patients who take drugs including nitrates, digoxin and serum sodium take higher risk of arrhythmia and are more needed for ICD, while patients who take ACE inhibitors are at lower risk to have arrhythmia and implant ICD. If the blood urea nitrogen (BUN) content is higher than 80 mg/dL, patients are more likely to suffer from arrhythmia, while patients with BUN less than 80 have lower hazard ratio thus are free from installing ICD. In addition, patients who own higher duration of PR interval, which is the time between the start of the P wave and the start of the R wave also in demand of ICD to be implanted.

6. Conclusions

Arrhythmia is a severe disease that threatens million people's lives per year. The goal of our report is to use survival analysis to identify characteristics of patients who are in barely need of implanting ICD to reduce the risk of heart attack. Our report provides several models and finally we choose the Cox regression model for

interpretation. However, for patients whose shock and death censored at different time, we failed to conclude these censored samples into our final model. Relating to congenital and acquired factors, there must be more correlation factors for high-risk patients in need of ICD. With other possible factors provided and more knowledge for dealing with censored data, further studies are worthy to be conducted.

7. References

[1] https://medschool.ucla.edu/cardiovascular-arrhythmia

[2] https://www.chf-solutions.com/heart-failure-classifications/

[3]https://www.sciencedirect.com/topics/neuroscience/qrs-complex#:~:text=Wide%20QRS%20complex%20tachycardia%20%28WCT%29%2C%20defined%20as%20heart,accurate%20diagnosis%20with%20initiation%20of%20appropriate%20therapy%20essential.

[4]https://www.mdapp.co/qtc-calculator-57/#:~:text=Abnormally%20high%20or%20low%20QTc%20Values%20lower%20than,the%20heart%20and%20can%20appear%20at%20any%20age.

[5]https://www.aurorahealthcare.org/services/heart-vascular/conditions/low-ejection-fraction

[6] Chen, Z., Sohal, M., Voigt, T., Sammut, E., Tobon-Gomez, C., Child, N., ... & O'Neill, M. (2015). Myocardial tissue characterization by cardiac magnetic resonance imaging using T1 mapping predicts ventricular arrhythmia in ischemic and non–ischemic cardiomyopathy patients with implantable cardioverter-defibrillators. *Heart Rhythm*, *12*(4), 792-801.

[7] Ng, A. C., Bertini, M., Borleffs, C. J. W., Delgado, V., Boersma, E., Piers, S. R., ... & Biffi, M. (2010). Predictors of death and occurrence of appropriate implantable defibrillator therapies in patients with ischemic cardiomyopathy. *The American journal of cardiology*, *106*(11), 1566-1573.

[8] Yildirir, A., Batur, M. K., & Oto, A. (2002). Hypertension and arrhythmia: blood pressure control and beyond. *Europace*, *4*(2), 175-182.

[9] Goonasekara, C. L., Balse, E., Hatem, S., Steele, D. F., & Fedida, D. (2010). Cholesterol and cardiac arrhythmias. *Expert review of cardiovascular therapy*, *8*(7), 965-979.

[10] Suzuki, S. (2011). " Cholesterol Paradox" in Atrial Fibrillation. *Circulation Journal*, 1110171434-1110171434.

[11] Nalliah, C. J., Sanders, P., Kottkamp, H., & Kalman, J. M. (2016). The role of obesity in atrial fibrillation. *European heart journal*, *37*(20), 1565-1572.