**SURVIVAL ANALYSIS PROJECT REPORT ON THE EFFECT OF AGE, BMI(BODY MASS INDEX), AGE AT DIAGNOSIS, SMOKING STATUS, SBP (SYSTOLIC BLOOD PRESSURE), DBP (DIASTOLIC BLOOD PRESSURE), ECG (ELECTROCARDIGRAM), CORONARY HEART DISEASE (CHD), ON THE SURVIVAL TIME OF 149 DIABETIC PATIENTS**

**BY**

**OGWURUMBA UCHECHUKWU ANTHONY**

**PANTHER NUMBER: 002-659-034**

1. **ABSTRACT**

This is a report about 149 diabetic patients. It was done to determine the effect of Age, BMI (Body Mass Index), Age at Diagnosis, smoking status, SBP (systolic blood pressure), DBP (Diastolic Blood Pressure), ECG (Electrocardigram) and Coronary Heart Disease (CHD) on the survival time of 149 diabetic patients.

The dataset was analyzed using R programming and for each of the levels of the categorical variables, the survival distribution was computed and while graphs were plotted and suggestions made. Also for each level, an appropriate estimator was obtain and confidence interval for the median of the survival curves were given and comments were made.

A single test of differences between the survival curves of the survival time of the diabetic patients was made where the log rank test method was used. This test method was imperative because the data are right skewed and [censored](https://en.wikipedia.org/wiki/Censoring_(statistics)).

The relative risks of survival for any pair of two levels of the categorical variable was computed. After which an appropriate model was built and using an appropriate estimator the confidence interval of the relative risks was computed and comments made.

Finally an appropriate parametric model was built for the dataset using Cox Proportional Hazard Regression Model and Aalen’s Additive Regression Model. While Cox Proportional Hazard Regression shows that only SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), ECG (Electrocardigram) have significant effect on survival time, Aalen’s Additive Regression Model shows that only SBP (Systolic Blood Pressure) and DBP (Diastolic Blood Pressure have significant effect on survival time. The former was chosen as the final model because of the robustness of Cox Proportional Hazard Regression Model.

1. **TABLE OF CONTENT**

Page

Front page…………………………………………………………………………………..…….1

Abstract……………………………………………………………………………………………2

Table of content…………………………………………………………………………………...3

Statement of objectives…………………………………………………………………………....4

Introduction……………………………………………………………………………………….5

Data Description…………………………………………………………………………………..6

Methods of Measurements………………………………………………………………………...7

Data Analysis……………………………………………………………………………………...8

Findings and Conclusions………………………………………………………………..………40

References………………………………………………………………………..………………41

.

1. **STATEMENT OF OBJECTIVES**

This survival analysis final project report is about 149 diabetic patients. It was carried out to;

* To determine the effect of Age, BMI (Body Mass Index), Age at Diagnosis, Smoking Status, SBP (systolic blood pressure), DBP (Diastolic Blood Pressure), ECG (Electrocardigram) and Coronary Heart Disease (CHD) on survival time 149 diabetic patients.
* To get their survival curves
* To obtain an appropriate estimator and confidence Interval for the median of the survival curves
* To test for difference between the survival curves
* To build an appropriate model for the dataset
* To obtain an estimator and confidence interval for the relative risks

1. **INTRODUCTION**

Survival timecan be defined broadly as the time to the occurrence of a given event (Lee et al., 1988). This event can be the development of a disease, response to a treatment, relapse, or death. Therefore, survival time can be tumor-free time, the time from the start of treatment to response, length of remission, and time to death.

Survival data can include survival time, response to a given treatment, and patient characteristics related to response, survival, and the development of a disease. The study of survival data has focused on predicting the probability of response, survival, or mean lifetime, comparing the survival distributions of experimental animals or of human patients and the identification of risk and/or prognostic factors related to response, survival, and the development of a disease.

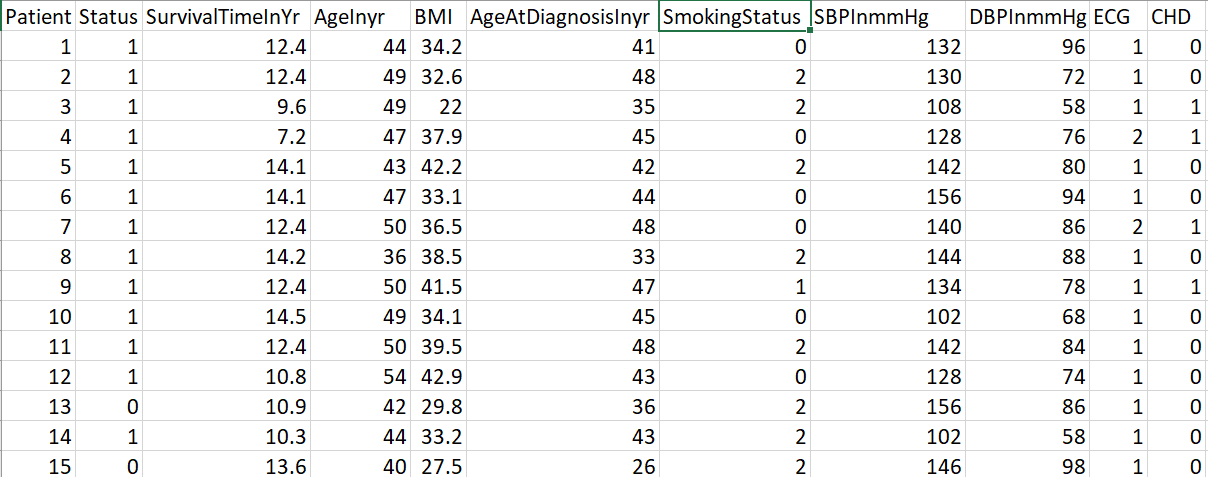
The endpoint of interest is the disease-free survival time, which is the time in years until death, relapse, or the end of the study. The focus is on the response variable, the censoring indicator (status of 0 indicates an event time, and a status of 1 indicates a censored time) and the categorical variable which are SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), Smoking Status, Coronary Heart Disease (CHD), ECG (ElectroCardiGram). Also Age, Age at Diagnosis, SBP (Systolic Blood Pressure) and DBP (Diastolic Blood Pressure) were recoded as categorical variables. In each of the levels of the categorical variables, the survival distribution is computed and their graph plotted and comments made.

1. **DATA DESCRIPTION**

The data that is used in this project come from a subset of data from Lee et al., 1988. It is a data where 149 diabetic patients were followed for 17 years. It gives the survival time from baseline examination, survival status (observed = 1 or censored = 0) and several potential prognostic factors at baseline: age, body mass index (BMI), age at diagnosis of diabetes, smoking status(0 for No, 1 for ex-smoker and 2 for current smoker), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Electrocardiogram reading (ECG) (1 for normal, 2 for borderline and 3 for abnormal), and whether the patient had any Coronary Heart Disease (CHD) (0 for No, 1 for Yes).

The below variables were recoded into categories;

* Age in years : < 50, above 50
* Age at Diagnosis: < 50, above 50
* Systolic Blood pressure (SBP) in mmHg : 119 for low, =120 for Normal, >120 for Abnormal
* Diastolic Blood pressure (DBP): <79 for low, =80 for Normal, >80 or Abnormal



……and so on till 149 patients

Source: Lee et al., 1988

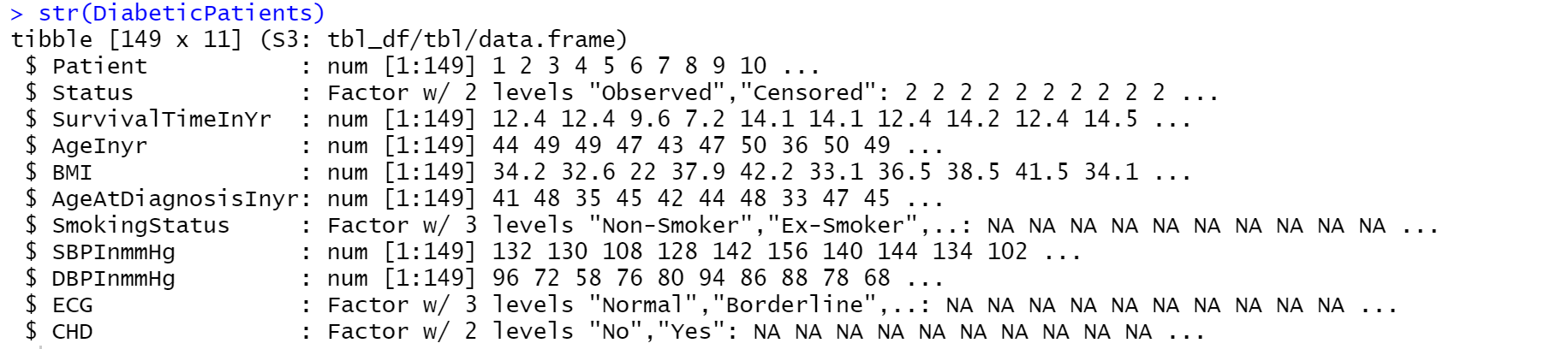
1. **METHODS OF MEASUREMENTS**

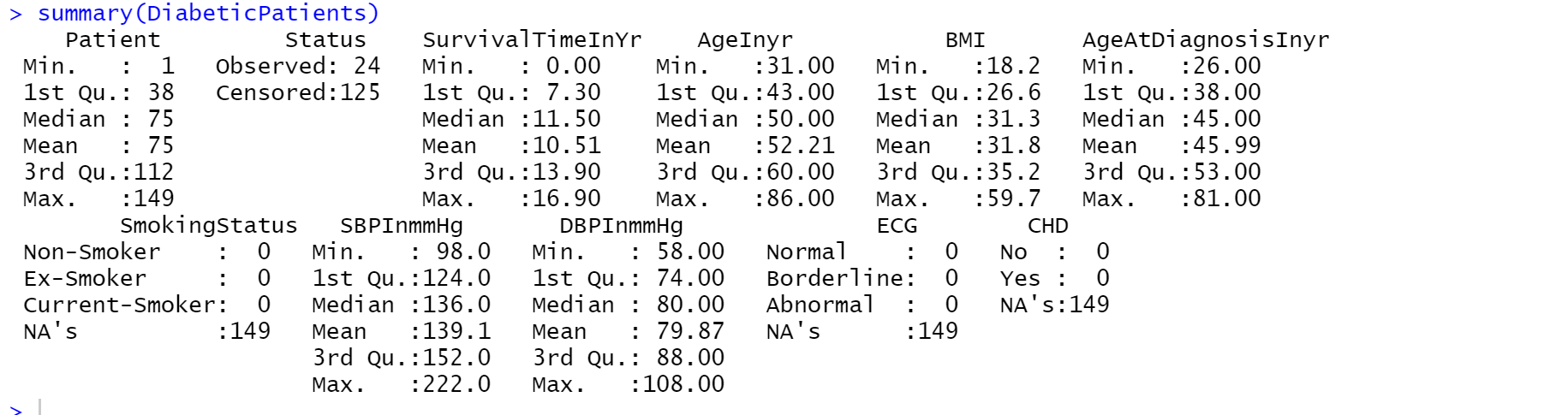
A qualitative method of measurement was used. Using the R programming package a detailed descriptive analysis using correlation analysis for the quantitative data and box plot for the categorical data was carried out to get a first impression of how the data look like. A survival analysis was carried out using Log Rant test, Pairwise Test. Then using the Cox Hazard Proportional Model and Aalen’s Additive Regression Model a test of significance of the variables was carried out after which they were used to build an appropriate model to the dataset.

1. **DATA ANALYSIS**

**7.1: Descriptive Analysis Of The Variables In The Dataset**

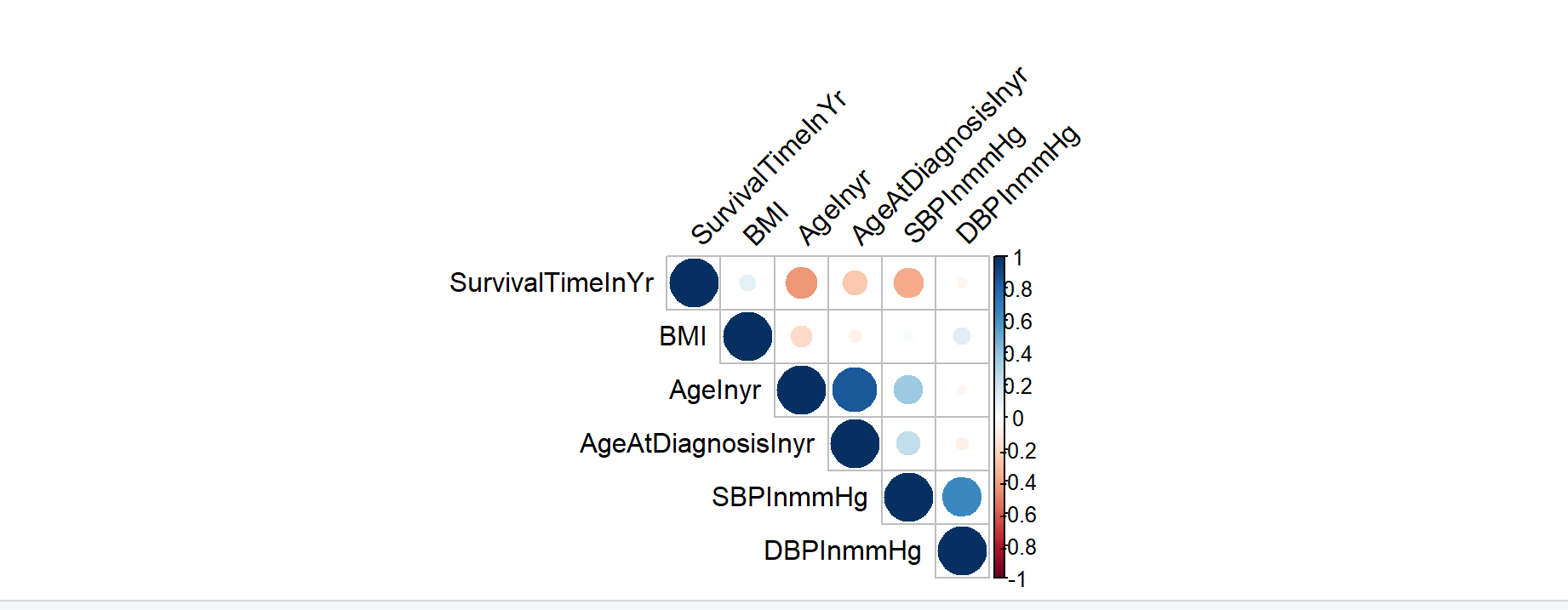
A detailed descriptive analysis of the dataset was performed using R and below are the result from the analysis;





**Comment:** From the above the mean of Age(yr), Body Mass Index(BMI), Age at Diagnosis(yr), Systolic Blood Pressure(SBP) , Diastolic Blood Pressure(DBP) haver mean of 52.2, 31.8, 45.9, 139.1 and 79.9 respectively.

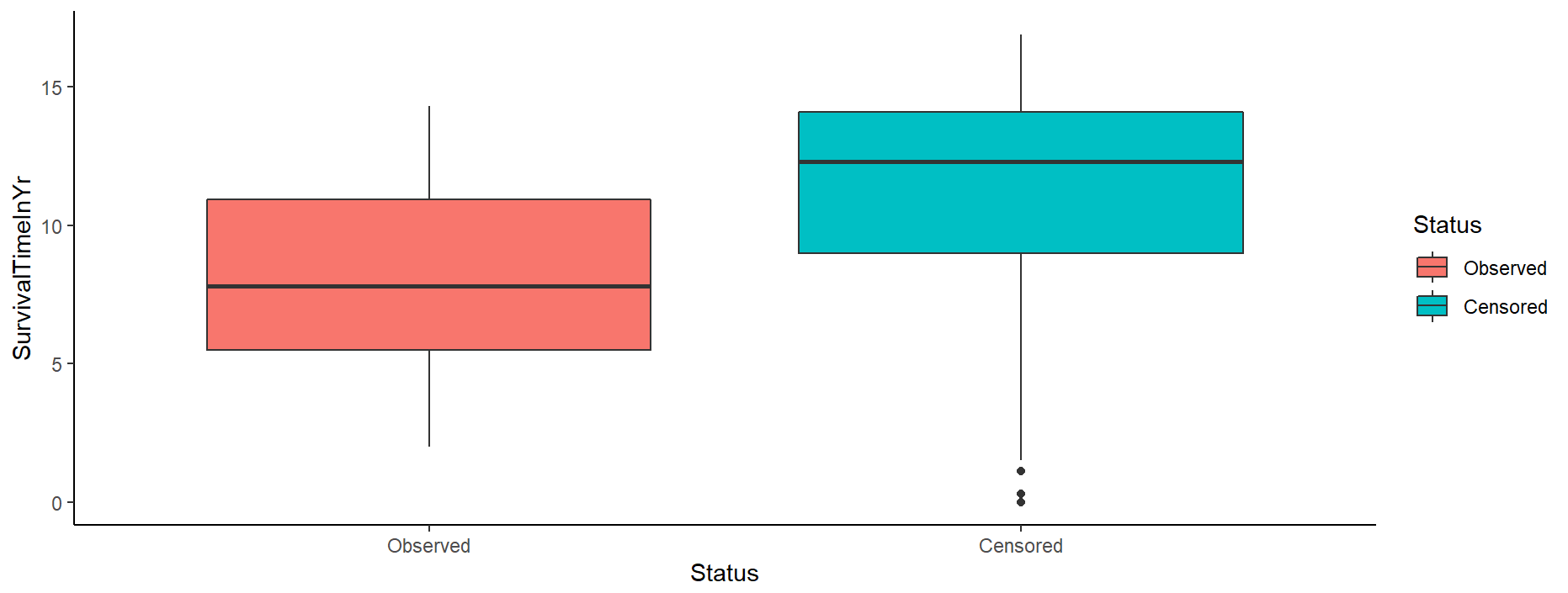
**7.1.1: Corrleation Of The The Surviavl Time Against The Quantitative Variables**



**Comment:** From the correlation diagram above Age in years is more correlated with survival time followed by SBP(in mmHg) and then Age at Diagnostic, DBP(in mmHg) and BMI in that order.

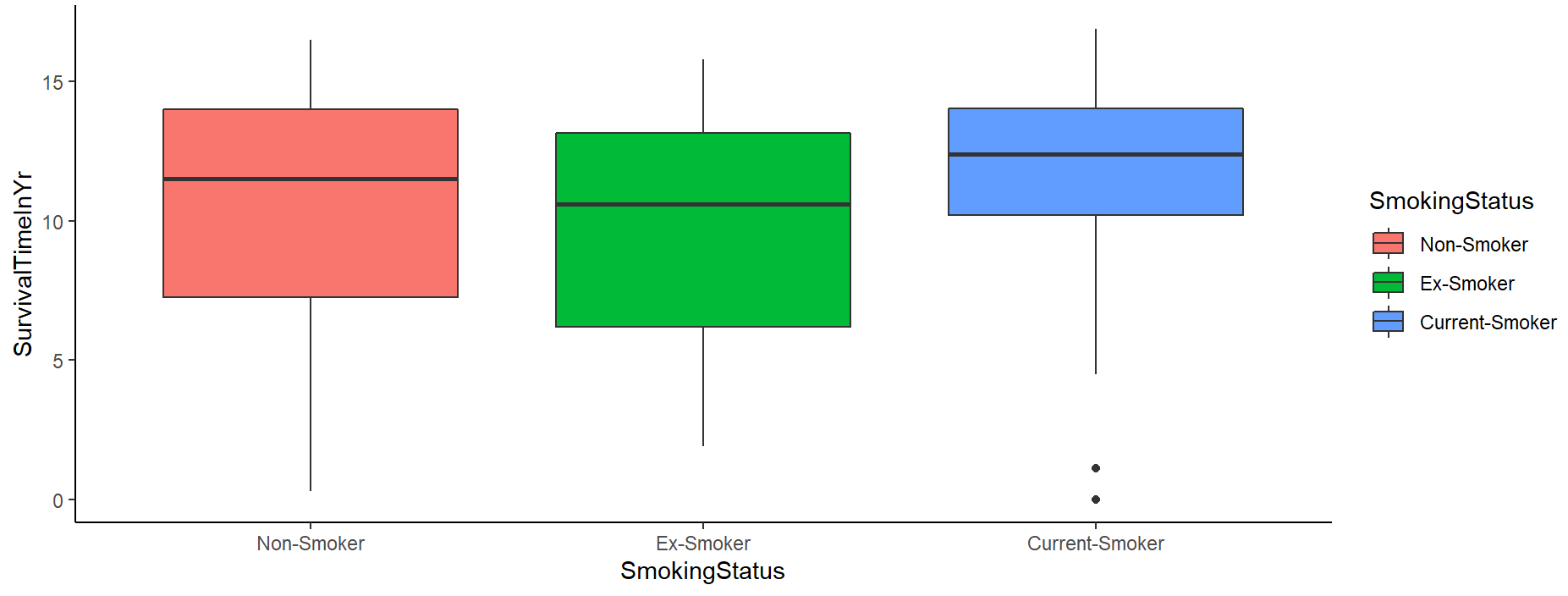
**7.1.2: Boxplot of the Categorical Variables**

Box Plot Of Survival Time using of Status group



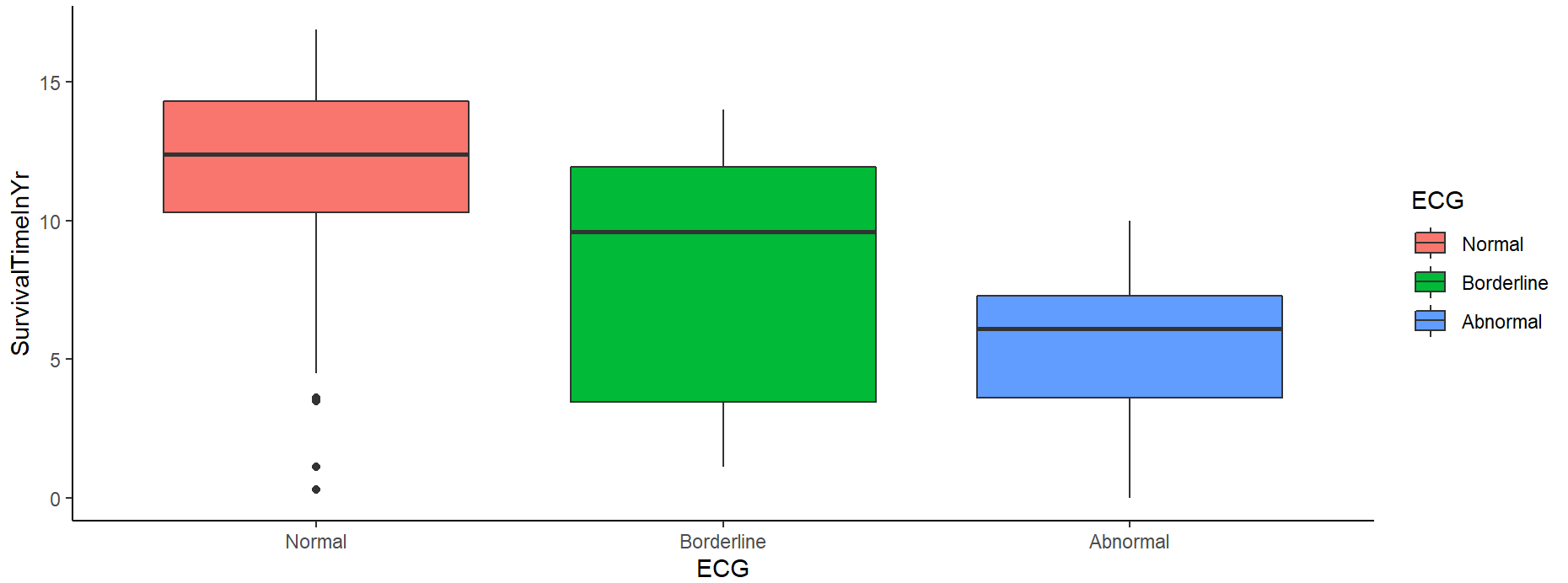
**Comment:** The survival time of most of the patients are censored

**7.1.3: Box Plot Of Survival Time Using Of Smoking Status Group**



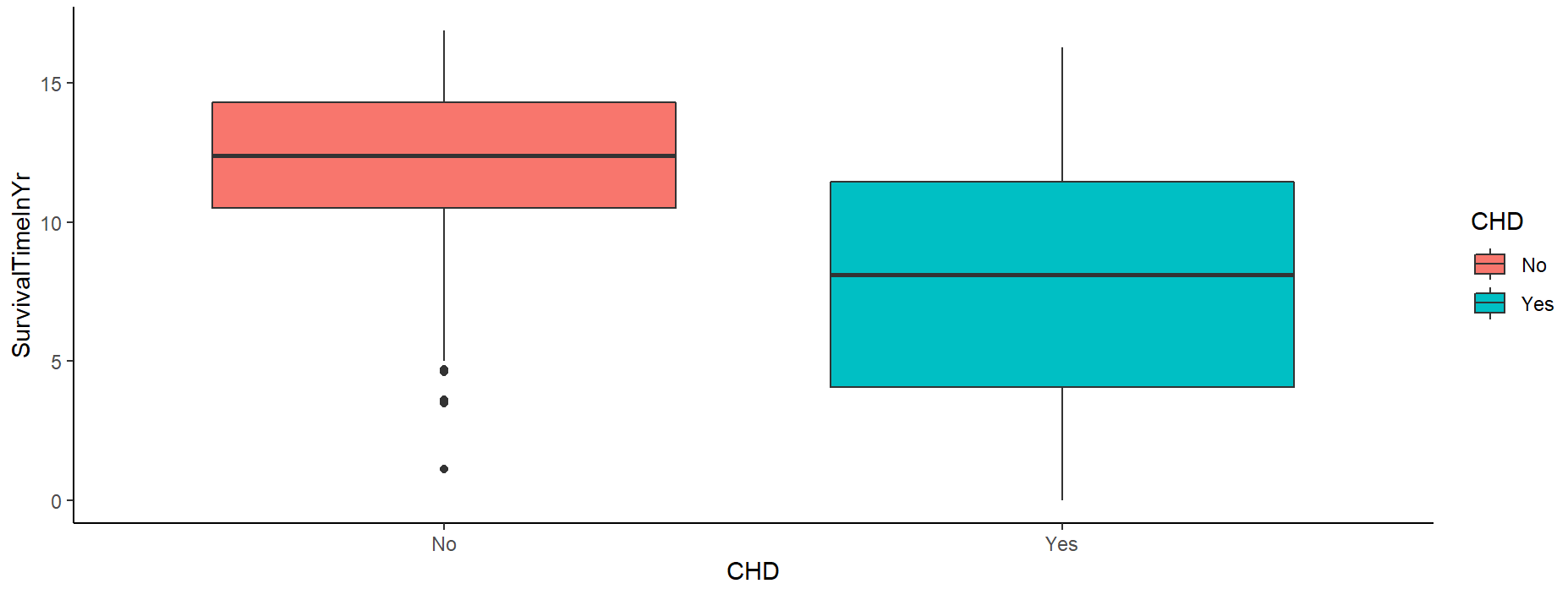
**Comment:** More patients are ex-smokers followed by Non-smokers

**7.1.4: Box Plot Of Survival Time Using Of Smoking Status Group**



**Comment:** More patients have borderline ECG reading

**7.1.5: Box Plot Of Survival Time Using Of CHD Group**

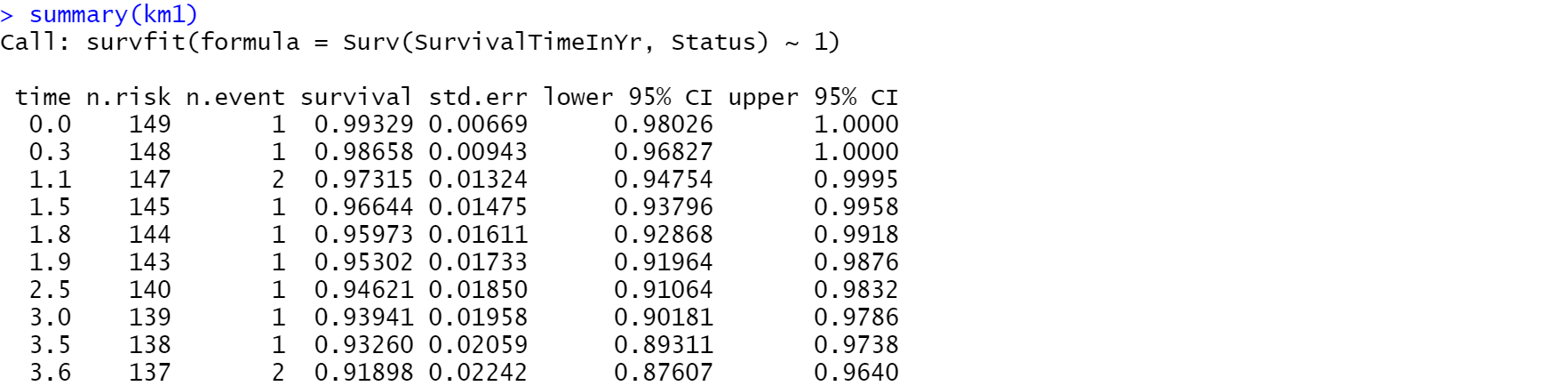


**Comment:** More patients have Chronic Heart Disease.

**7.2 The Kaplan Meier Plot**

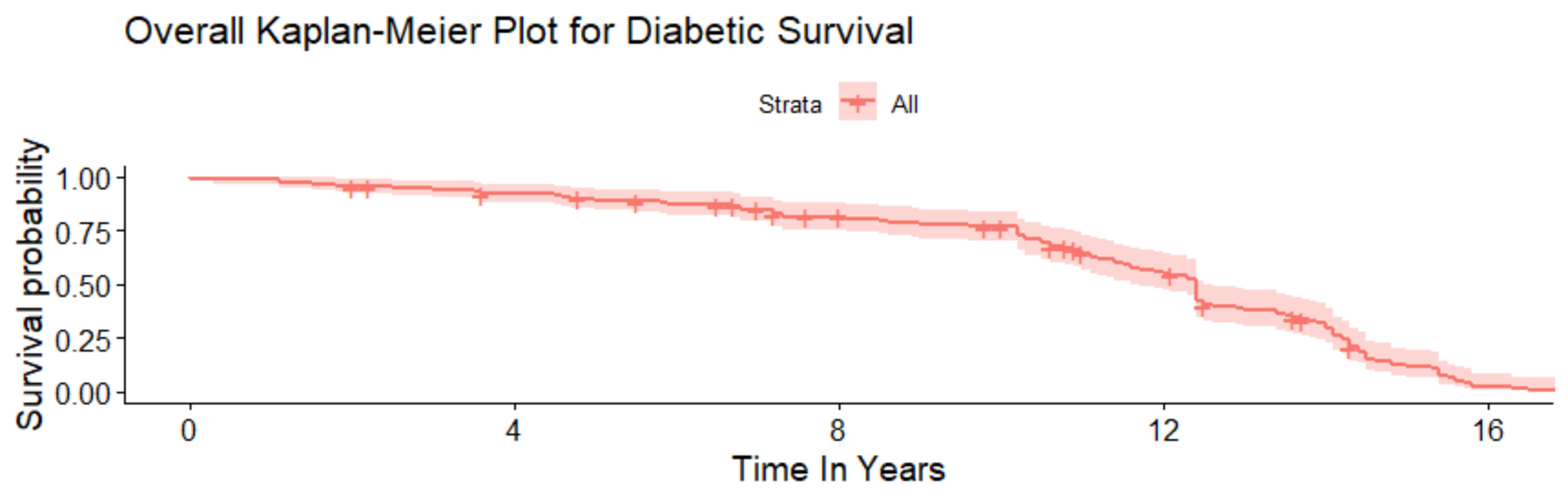
* The *x* axis is time, from zero (when observation began) to the last observed time point.
* The *y* axis is the proportion of subjects surviving. At time zero, 100% of the subjects are alive without an event.
* The solid line (similar to a staircase) shows the progression of event occurrences.
* A vertical drop indicates an event.

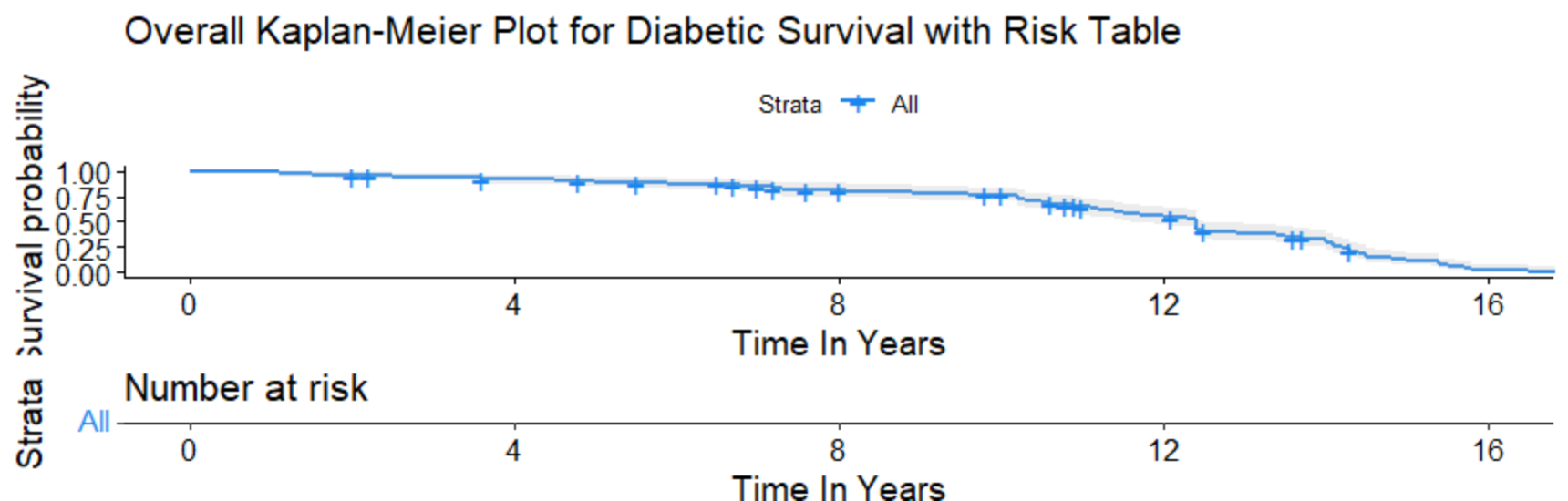
**7.2.1: The survival distribution using Kaplan Meier’s Estimate**



…upto 16.9 years

**7.2.2: Graph Of The Survival Distribution curve**

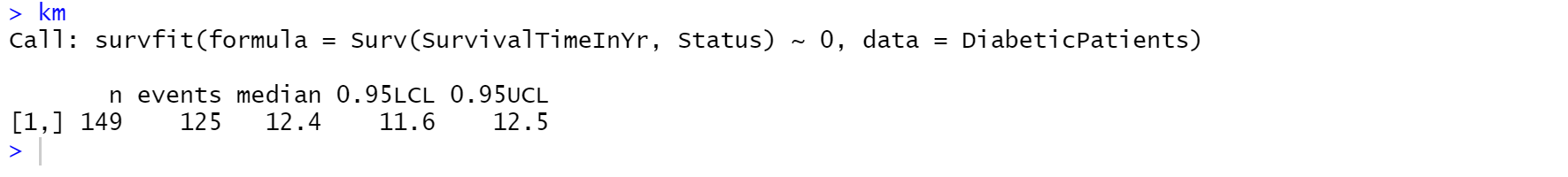


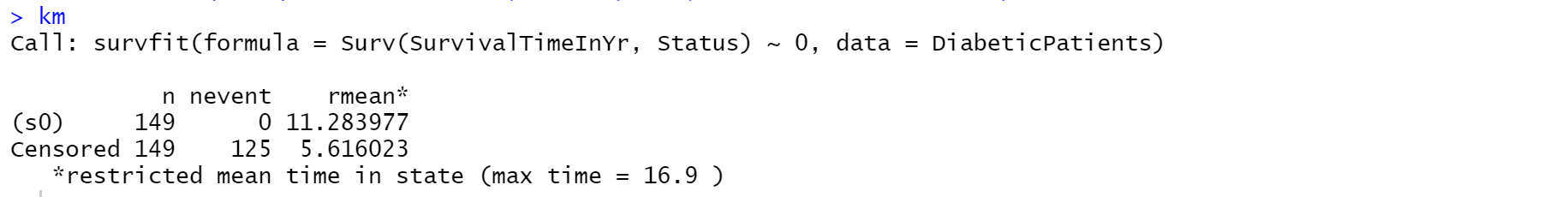


**Comment:**More than half of the patients with diabetics have a survival probability of more than 0.50. They last more than half of the 17 years period that the patients were followed.

**7.2.3: The 95% Confidence Interval For The Median Of The Survival Curves The Data Set**

Using the Kaplan Meier Estimates the confidence interval for the median of the survival curves using all the variables are:

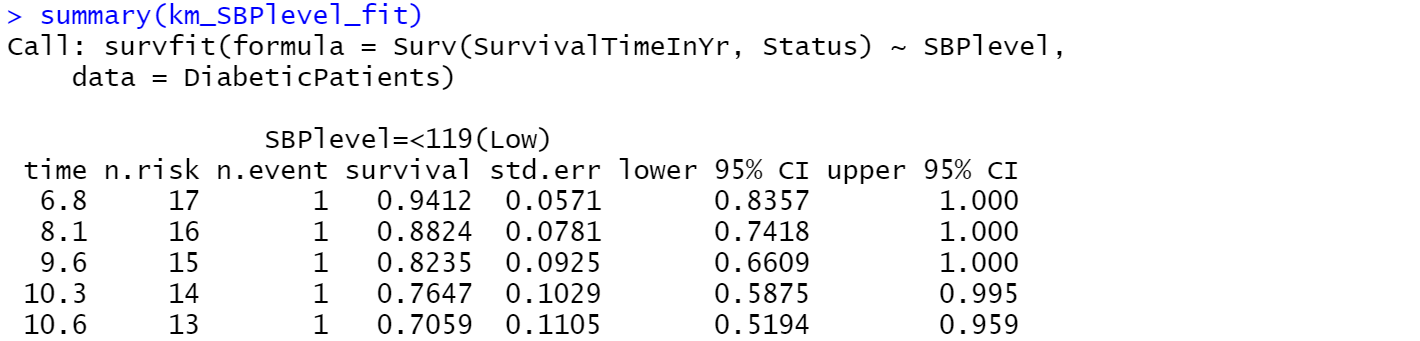
****

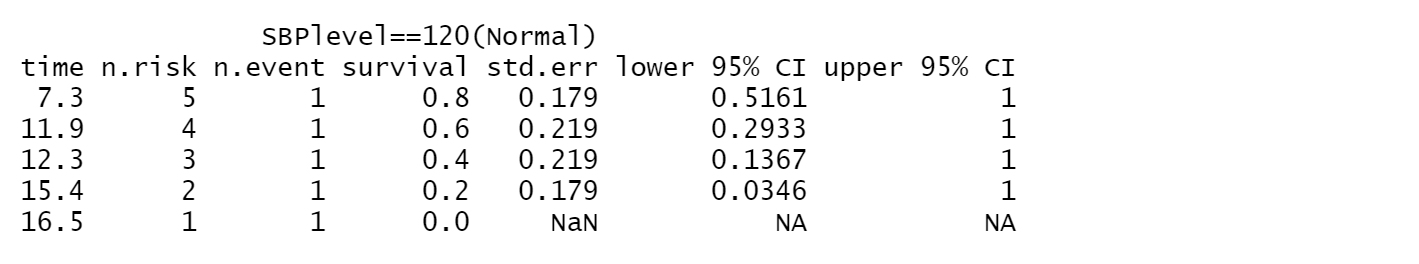
****

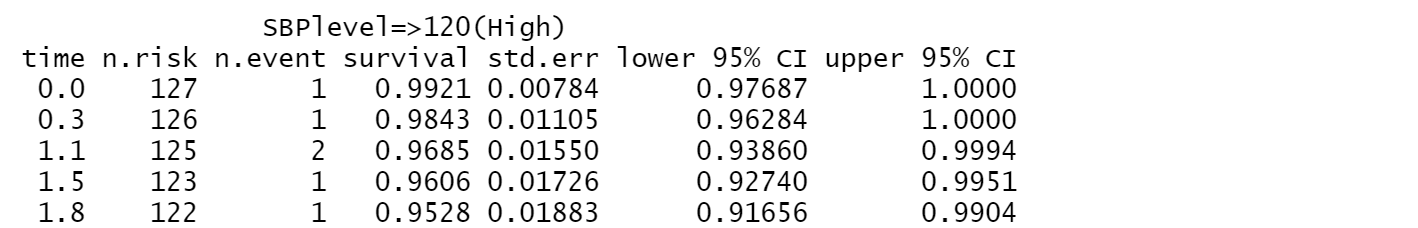
**Comment:**The median survival time for patients with diabetics is 12.4years with a 95% confidence interval of 11.6 – 12.5years .The maximum survival time of patients with diabetics is of 16.9 years.

**7.2.4: The Survival distribution of the SBP in mmHg**

Below is the distribution of SDP in mmHg at each levels.

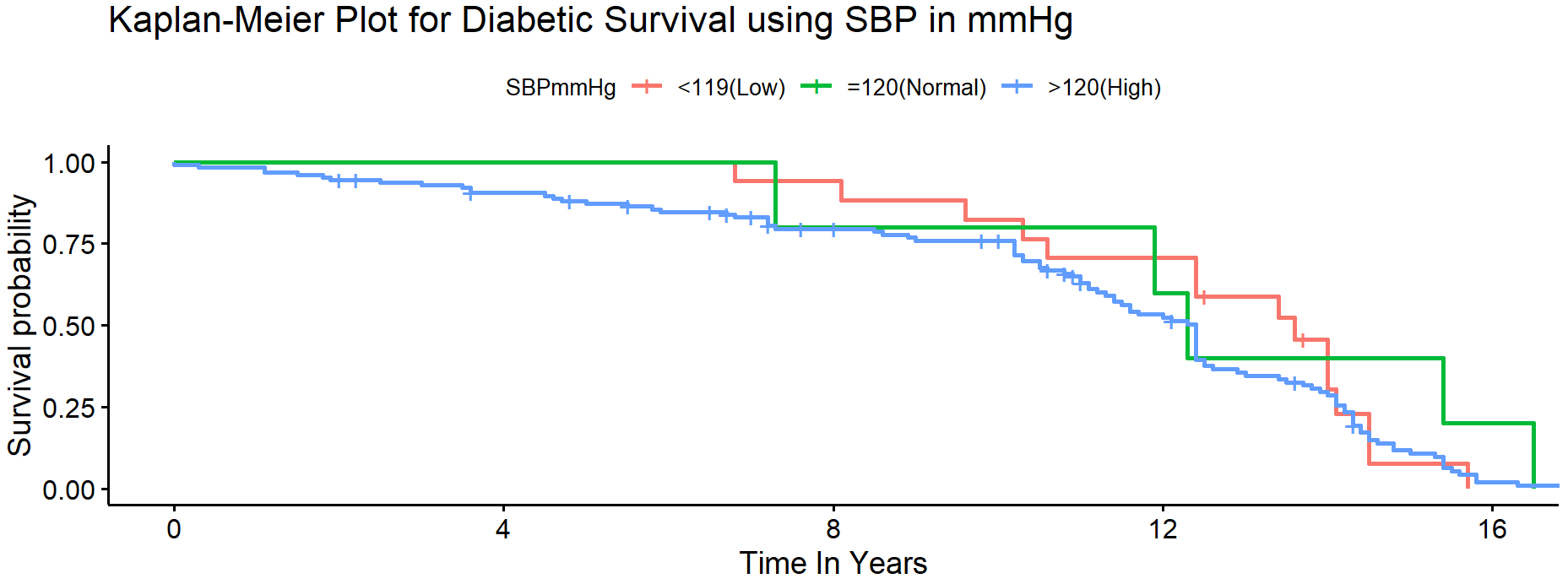
 ……and so on upto 15.7 years

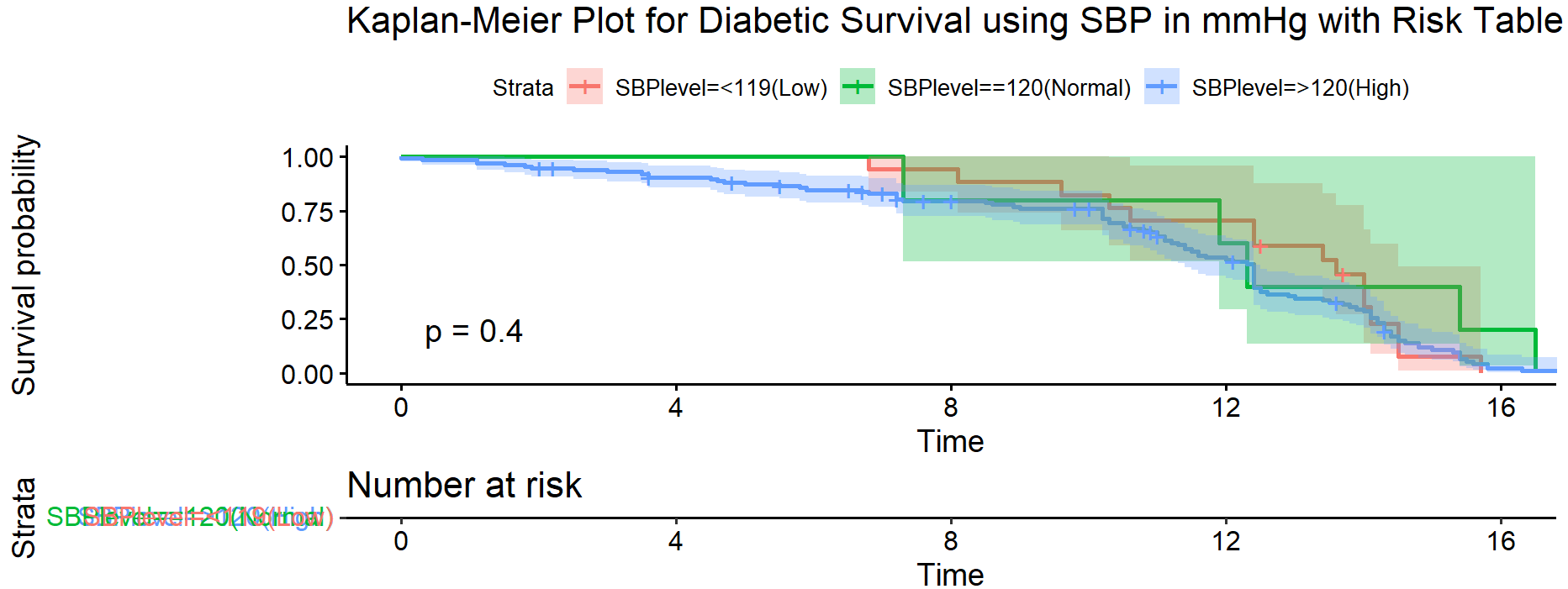




……and so on upto 16.9 years

**7.2.5: Graph Of The Survival Distribution Curve Of SBP In mmHg At Each Level**

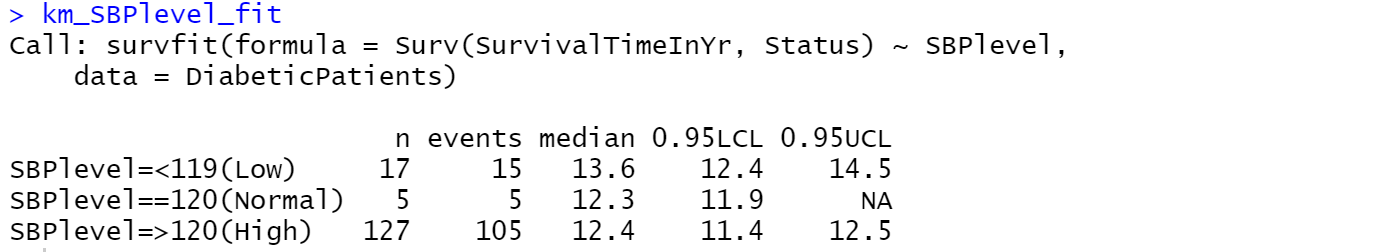




**Comment:** The above is the plots of the SDP at each level. It has levels which are red for <119(Low), green for =120Normal) and blue for >120(High). It can be seen from the distribution curve that the patients with normal SBP have higher probability to survive than the patients with low and high SDP.

**7.2.6: The 95% Confidence Interval For The Median Of The Survival Curves At Each Level Of SBP**

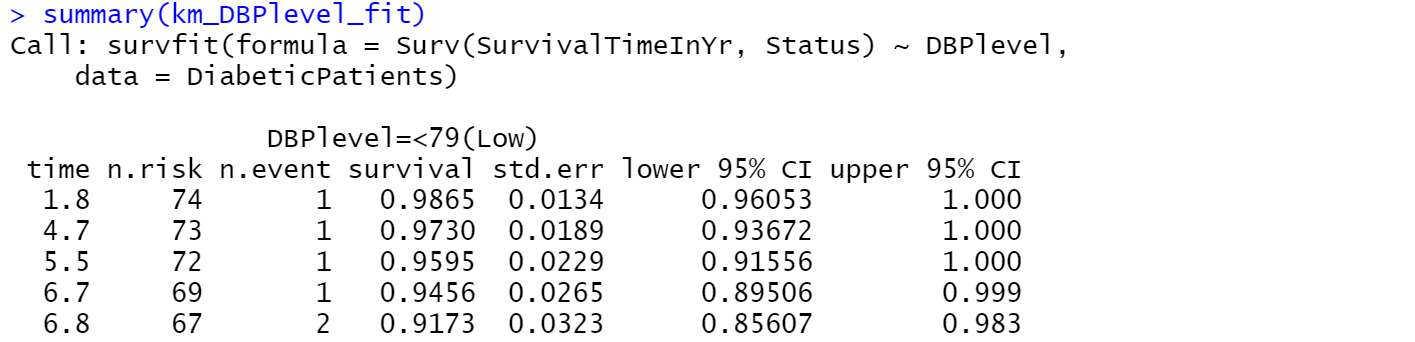
Using the Kaplan Meier Estimates the confidence interval for the median of the survival curves of SBP are:



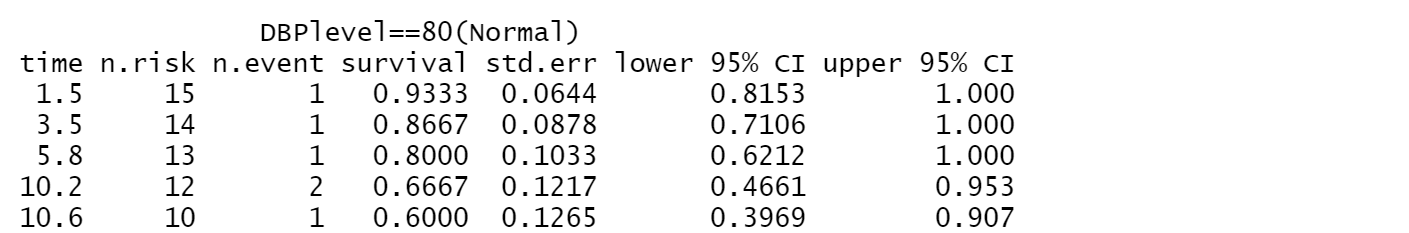
**Comment:**The median survival time of patients with a low Systolic Blood Pressure of less than 119 is 13.6years with a 95% confidence interval of 12.4 -14.5years . Those patients with normal Systolic Blood Pressure equal to 120 have median survival time of 12.3years with a 95% confidence interval of 11.9 years and above. Patients with high Systolic Blood Pressure greater than 120 have median survival time of 12.4years with a 95% confidence interval of 11.4-12.5years. This shows that patients with low Systolic Blood Pressure live longer.

**7.2.7: The Survival Distribution Of The DBP In mmHg**

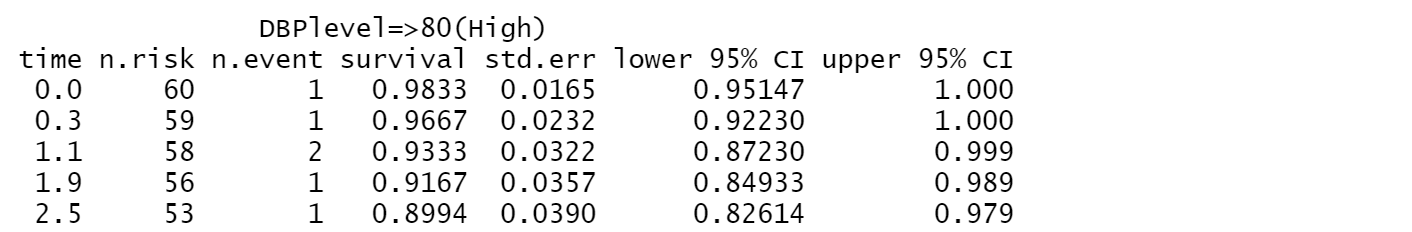
Below is the distribution of DBP in mmHg at each levels



……and so on upto 16.5 years

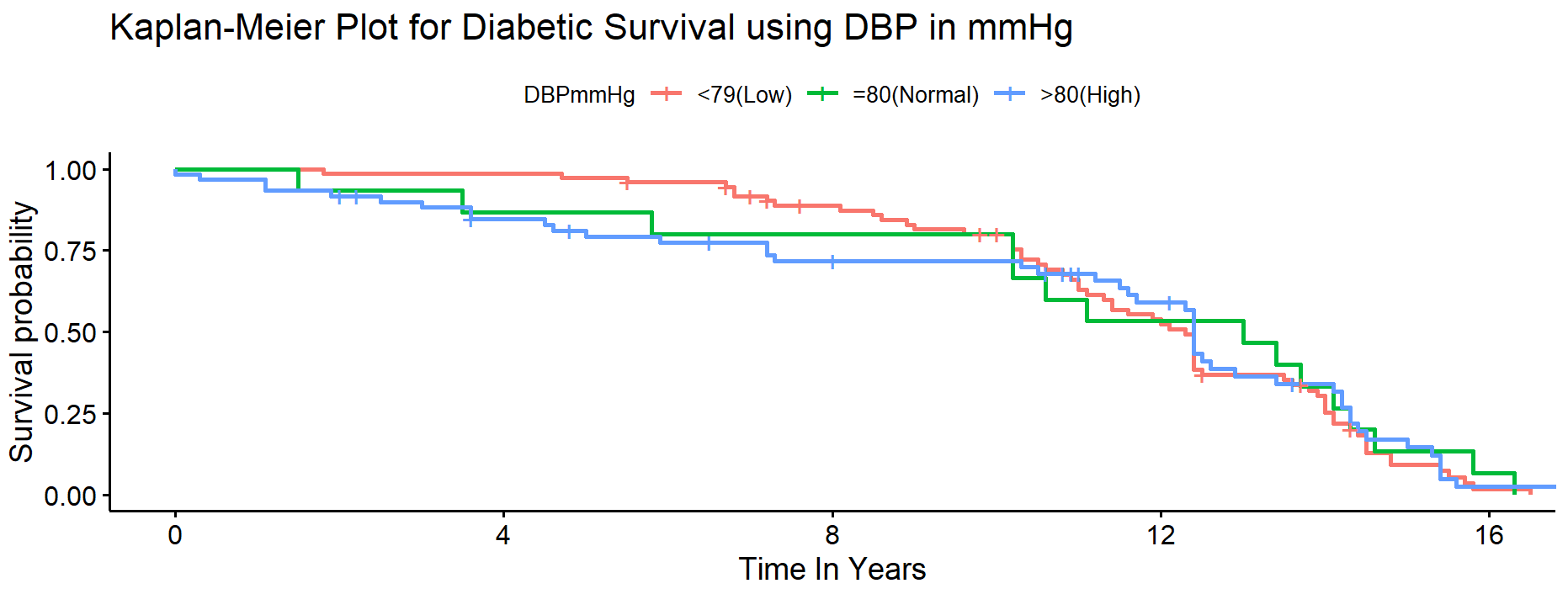


……and so on upto 16.3 years

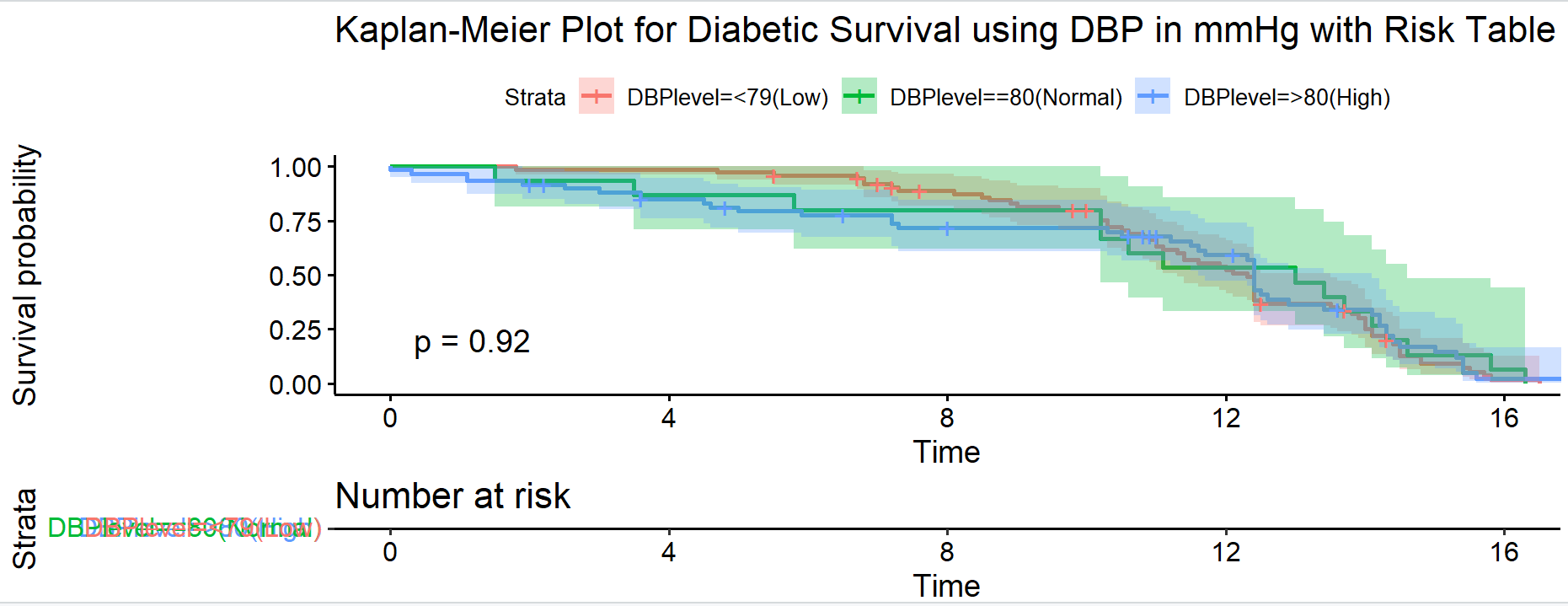


……and so on upto 16.9 years.

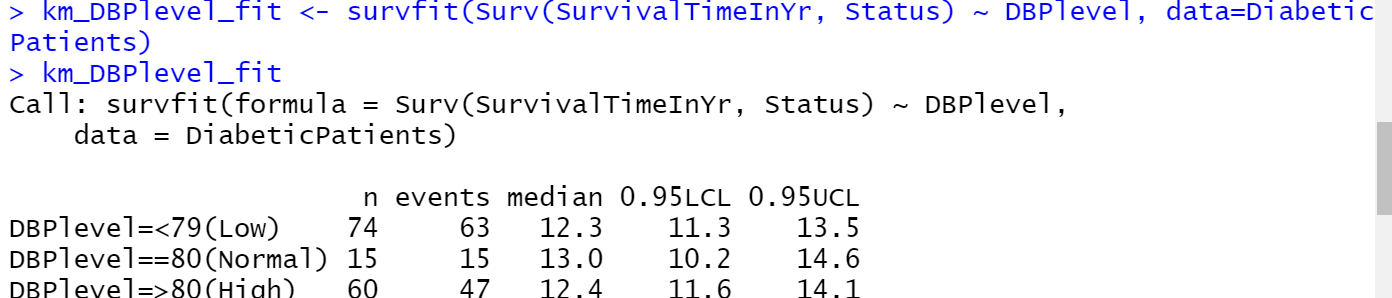
**7.2.8: Graph Of The Survival Distribution Of DBP At Each Level**



**Comment:**It can be seen from the distribution curve that patients with low DBP have higher survival probability than patients with Normal and High DBP.



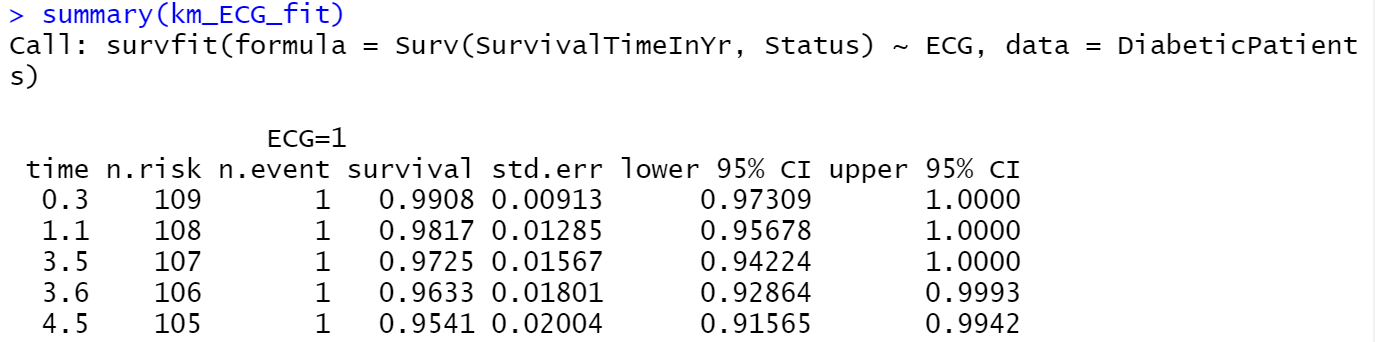
**7.2.9: The 95% confidence interval for the median of the survival curves using all DBP**

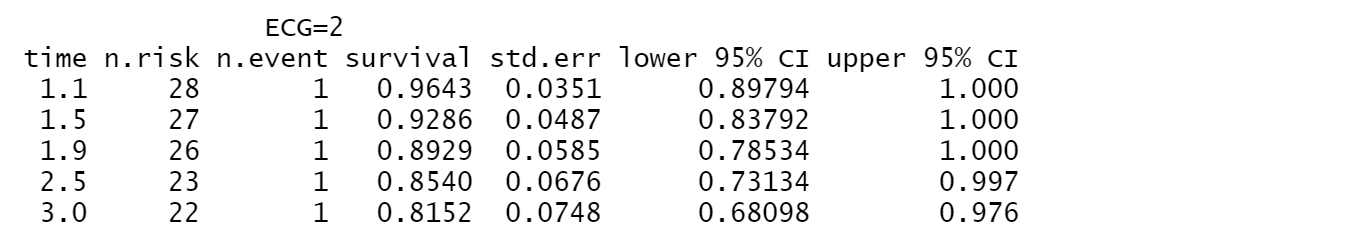


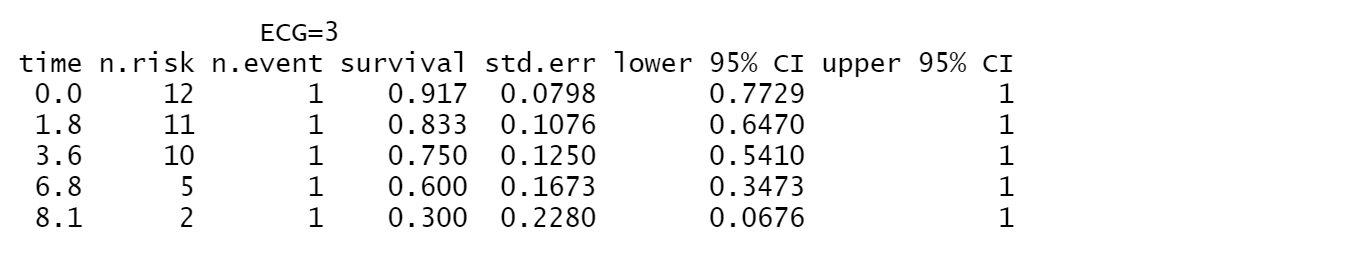
**Comment:**The median survival time of patients with a low Diastolic Blood Pressure of less than 79 is 12.3years with a 95% confidence interval of 11.3-13.5 years . Those patients with normal Diastolic Blood Pressure equal to 80 have median survival time of 13years with a 95% confidence interval of 10.2-14.6 years. Patients with high Diastolic Blood Pressure greater than 80 have median survival time of 12.4years with a 95% confidence interval of 11.6-14.1years. This shows that patients with normal Diastolic Blood Pressure live longer.

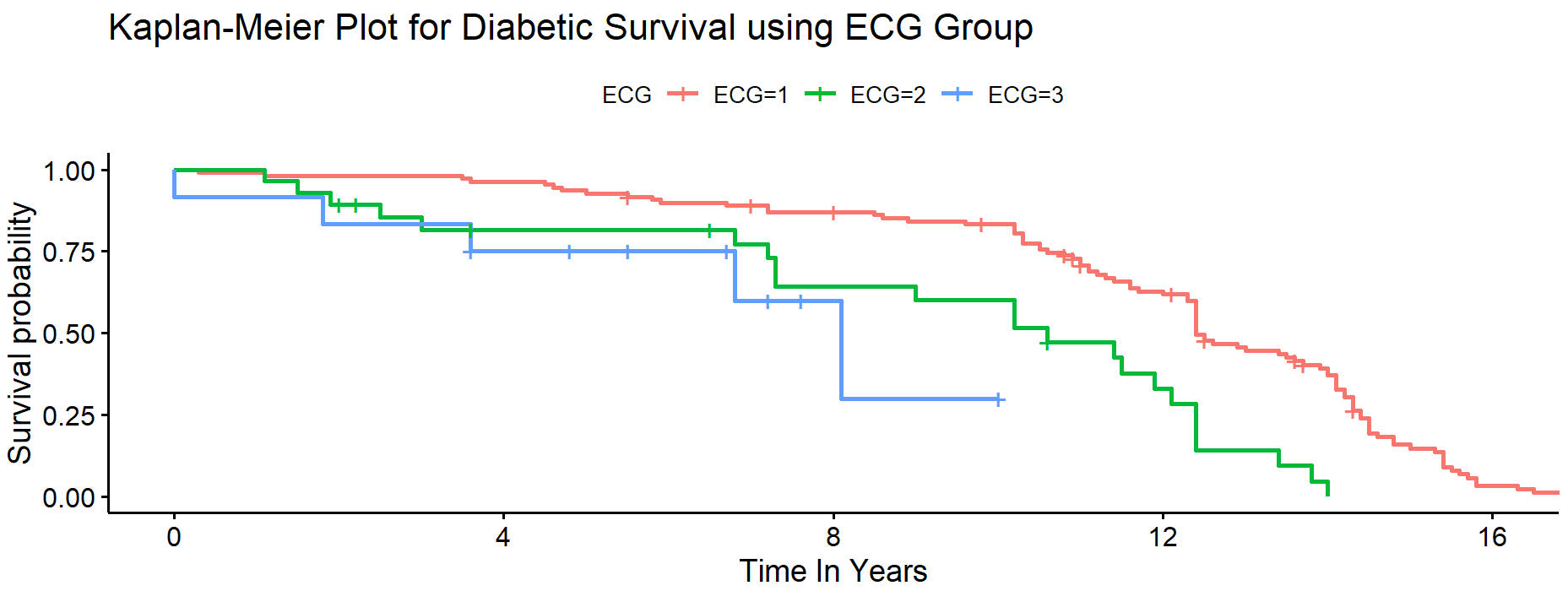
**7.2.10: The Survival Distribution of the ECG**

Below is the distribution of ECG at each levels.

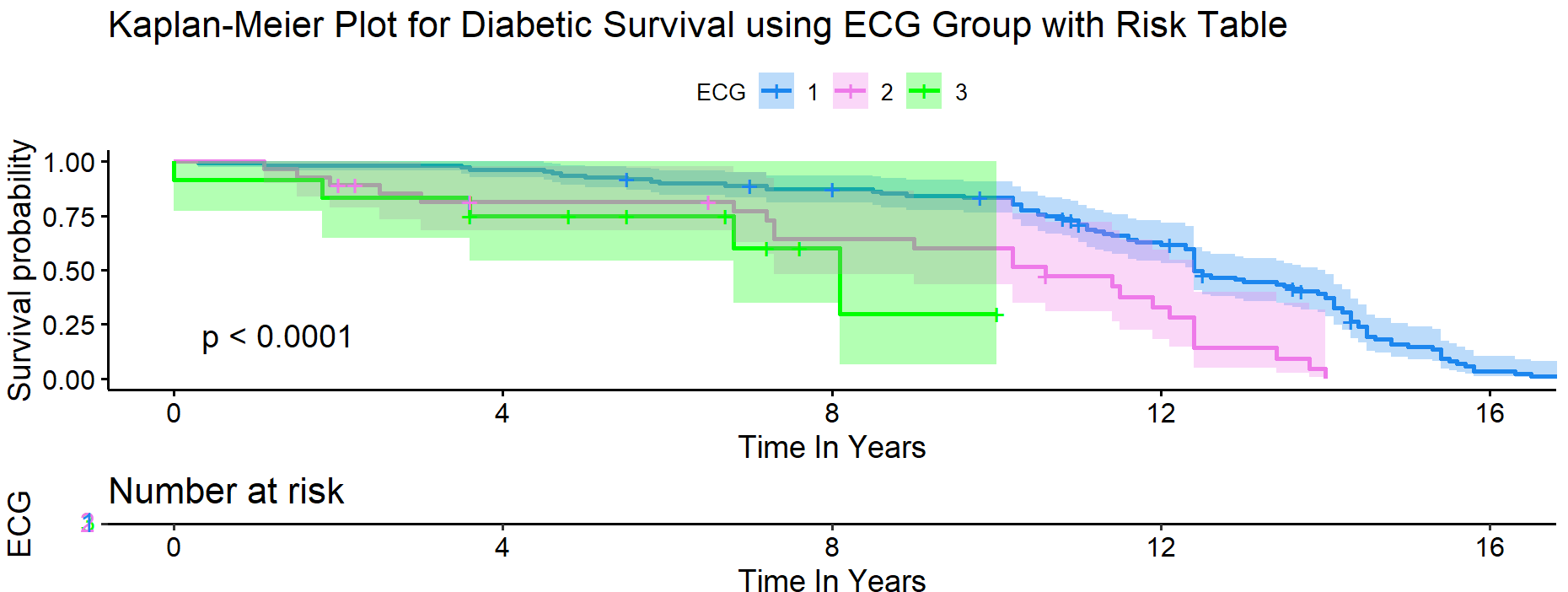
……and so on upto 16.9 years

……and so on upto 14 years



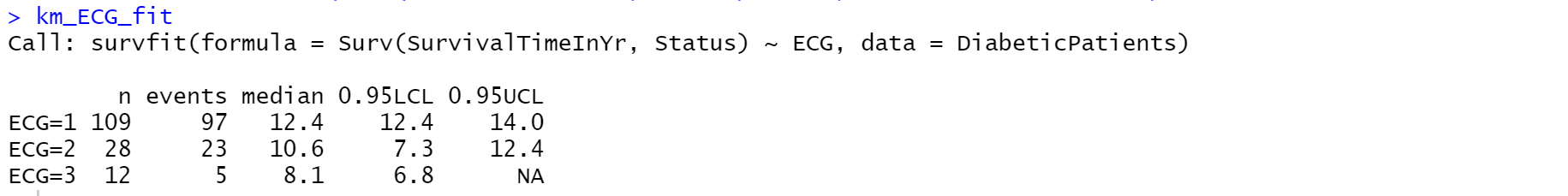
**7.2.11: Graph Of The Survival Distribution Of ECG At Each Level** 

**Comment:**It can be seen from the distribution curve that patients with Normal(1) ECG reading have higher survival times than patients with Borderline(2) and Abnormal(3) readings.



**Comment:**The above are plots of ECG at each levels, The levels are 1 for Normal (blue), 2 for Boarderline(pink) and 3 for Abnormal(green). It can be seen from the distribution curve that patients with Normal(1) ECG reading have higher survival times followed by patients with Borderline(2) and Abnormal(3) readings. The p-value of 0.0001 shows that ECG is significant to survival times.

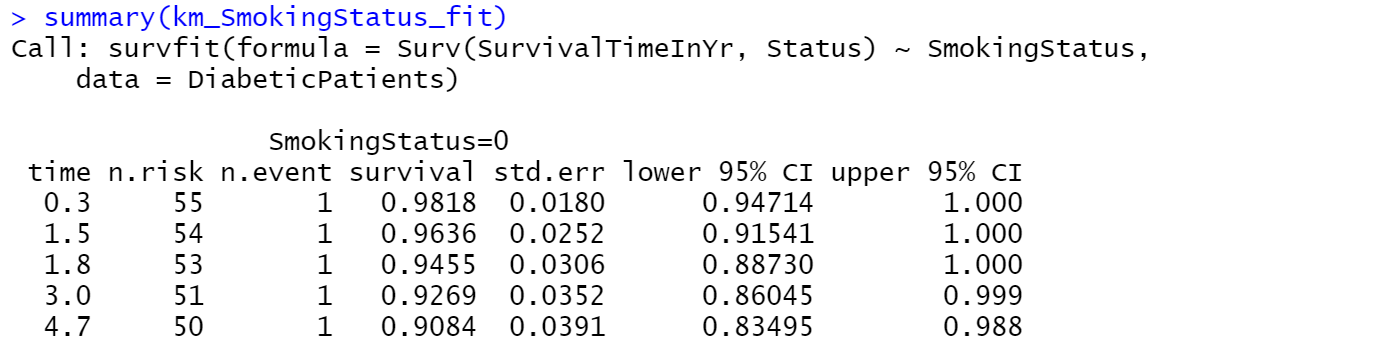
**7.2.12: The 95% confidence interval for the median of the survival curves using all ECG**

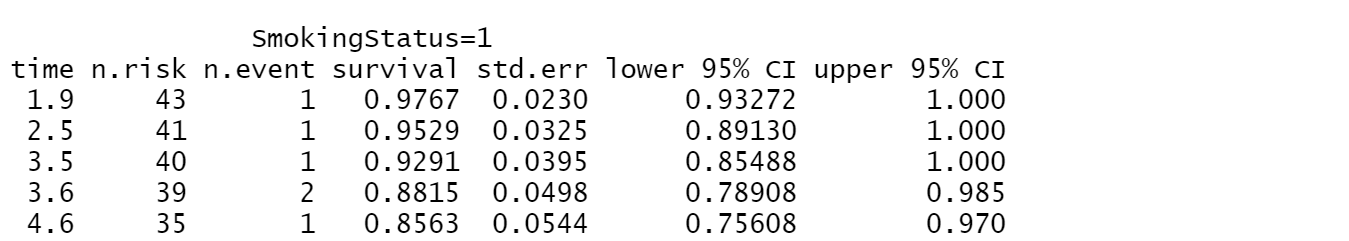


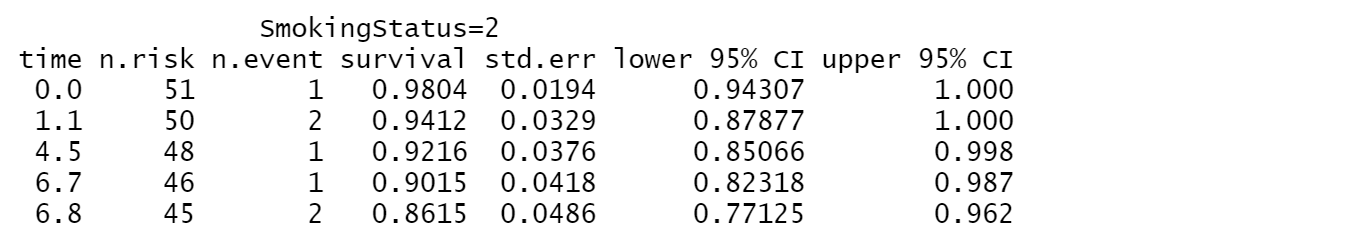
**Comment:**The median survival time of patients with Normal Electrocadiogram readings is 12.4years with a 95% confidence interval of 12.4- 14 years . Those patients with borderline Electrocadiogram readings have a mean survival time of 10.6 years with a 95% confidence interval of 7.3- 12.4 years. That of patients with Abnormal Electrocardiogram reading have a mean survival time of 8.1years and a 95% confidence interval of 6.8 and above. This shows that ECG reading has a lot of effect on survival time.

**7.2.13: The Survival Distribution Of The Smoking Status**

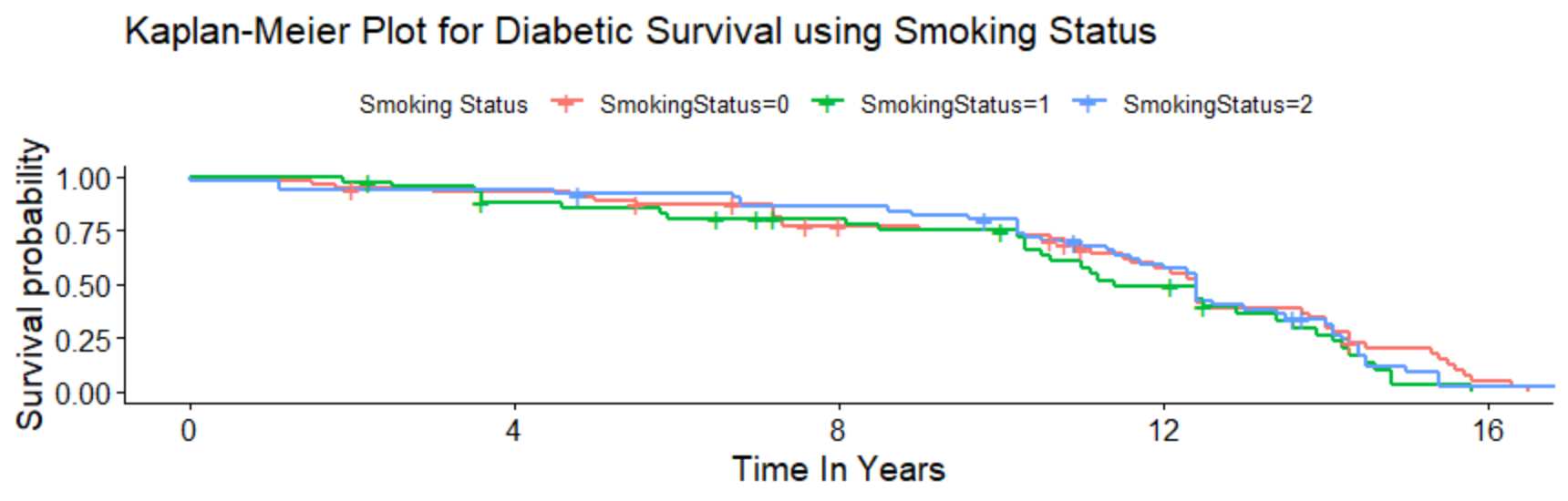
Below is the distribution of smoking Status at each levels.

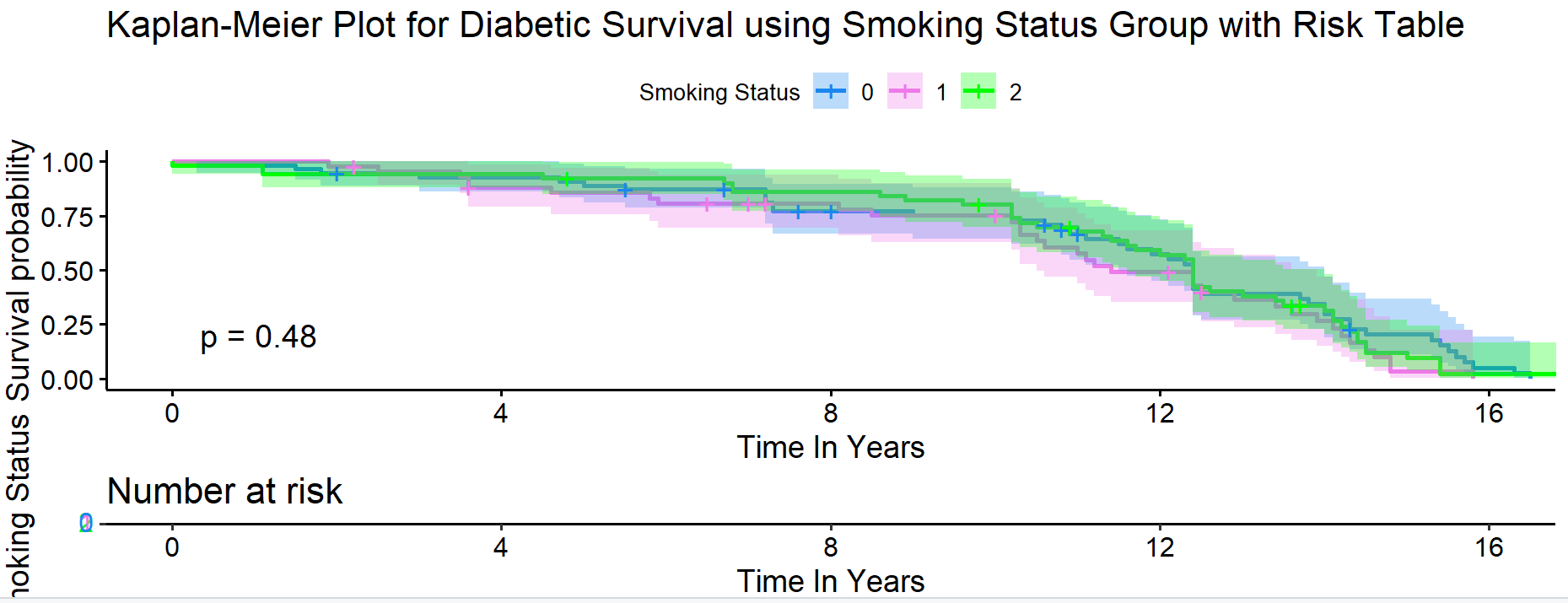
……and so on upto 16.5 years

……and so on upto 15.8 years

……and so on upto 16.9 years

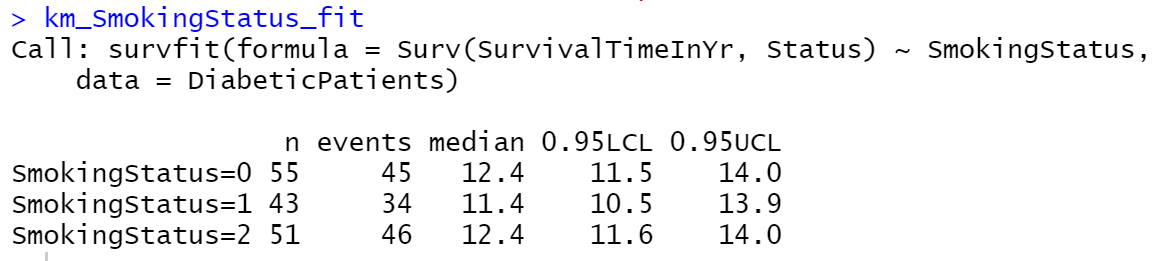
**7.2.14: Graph Of The Survival Distribution Curve Smoking Status At Each Level**





**Comments:**The above is the plots of the smoking status at each level. It has levels which are 0 (pink color) for non-smoker, 1(green color) for ex-smoker and 2(blue color) for current smokers. It can be seen from the distribution curve that even though current smokers experience more deaths(event) than non-smokers and ex-smokers they tend to have more survival time. It also suggests that smoking status have no significant effect on survival times.

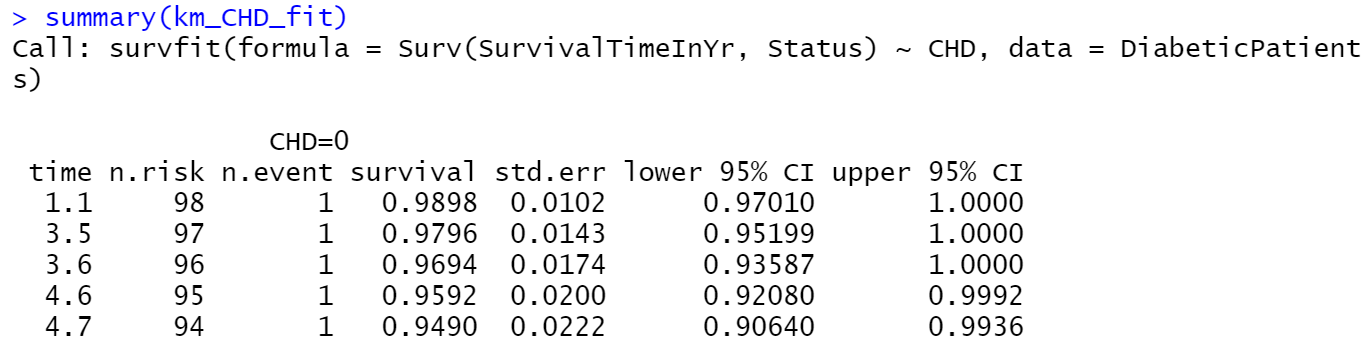
**7.2.15: The 95% Confidence Interval For The Median Of The Survival Curves Using All Smoking Status**

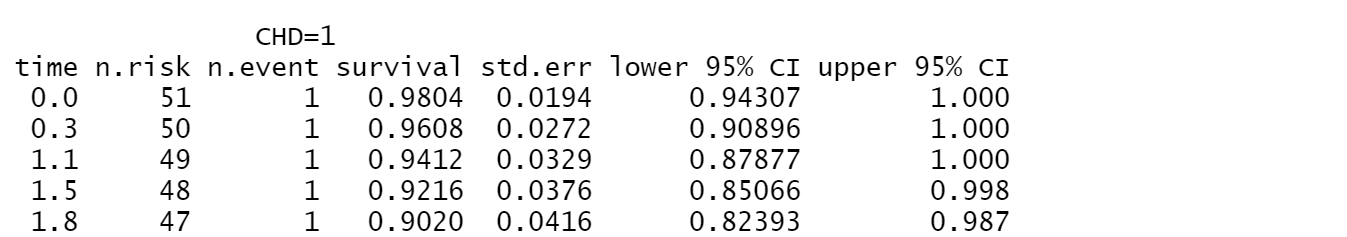


**Comment:**The median survival time of patients that do not smoke is 12.4years with a 95% confidence interval of 11.5-14.0 years while ex-smoker patients have a median survival time 11.4 years and a 95% confidence interval of 10.5-13.9 years. Patients that are smokers have a median survival time of 12.4years with a 95% confidence interval of 11.6-14.0 years. This is also showing that smoking status has no effect on survival time of diabetic patients.

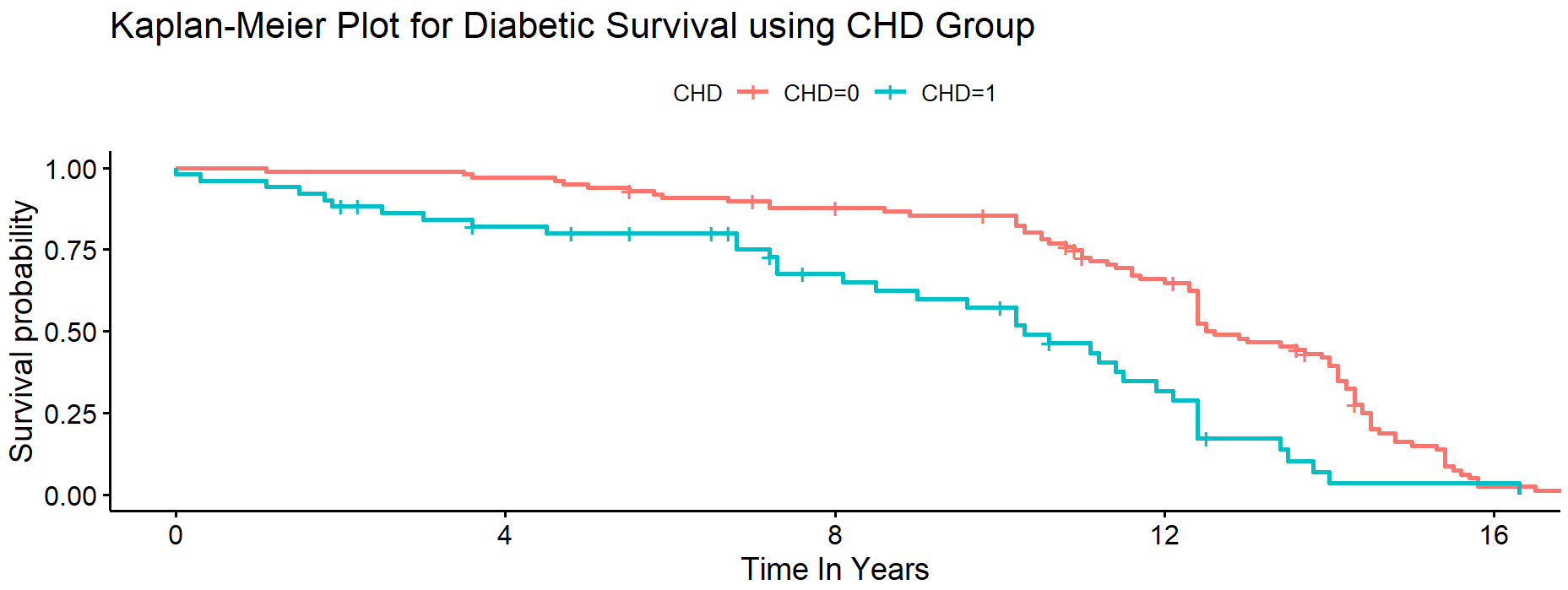
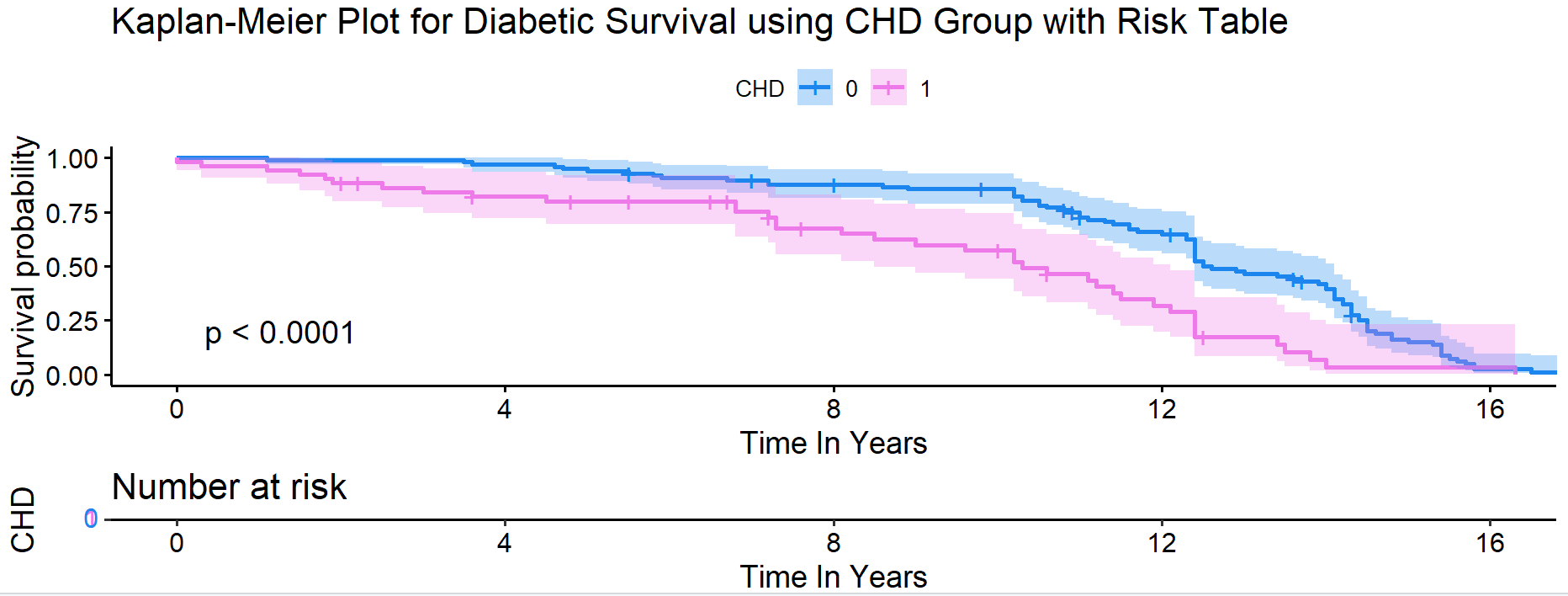
**7.2.16: The Survival Distribution Of CHD group**

Below is the distribution of CHD at each levels.

……and so on upto 16.9 years

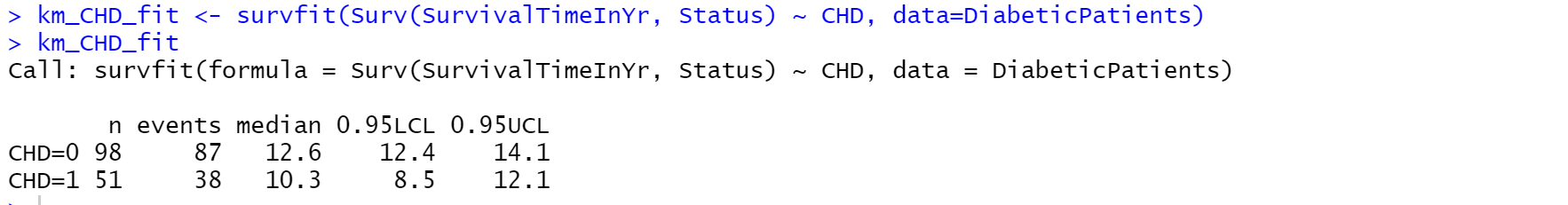
……and so on upto 16.3 years

**7.2.17: Graph Of The Survival Distribution Curve CHD At Each Level**

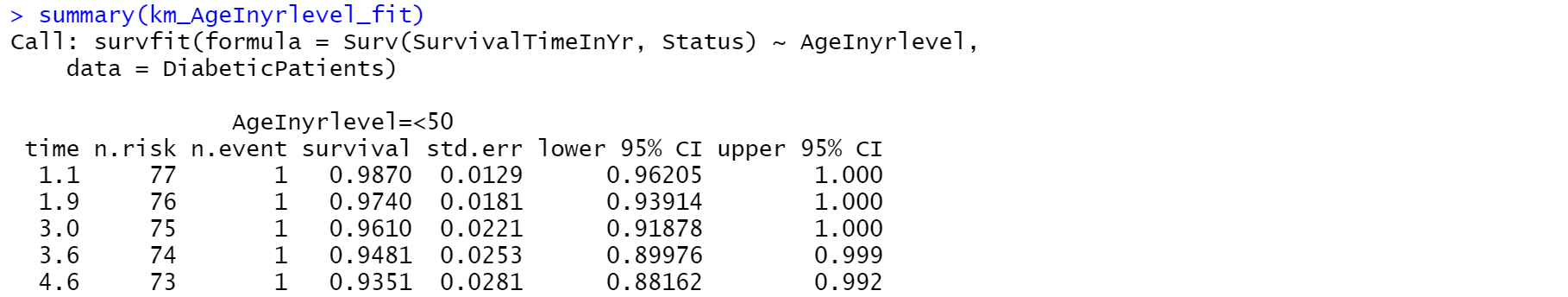
**Comments:** The above are plots of CHD at each levels. The levels are 0 for No (blue) to Coronary Heart Disease and 1 for Yes(pink). It can be seen from the distribution curve that those diabetic patients that do not have Coronary Heart Disease have higher survival rate than those that have it. It has a p value is 0.0001 which makes it significant to survival time.

**7.2.18: The 95% Confidence Interval For The Median Of The Survival Curves Using CHD**

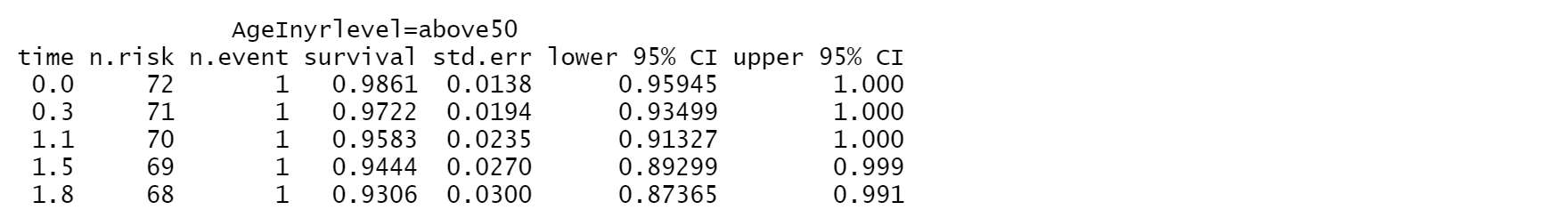


**Comment:**The median survival time of patients with no Coronary Heart Disease is 12.6years with a 95% confidence interval of 12.4- 14.1 years . Patients with Coronary Heart Disease have a median survival time 10.3 years and a 95% confidence interval of 8.5-12.1 years. It shows that patients with no Coronary Heart Disease have more survival probability.

**7.2.19: The Survival Distribution Of Age (In Yrs) At Each Level**

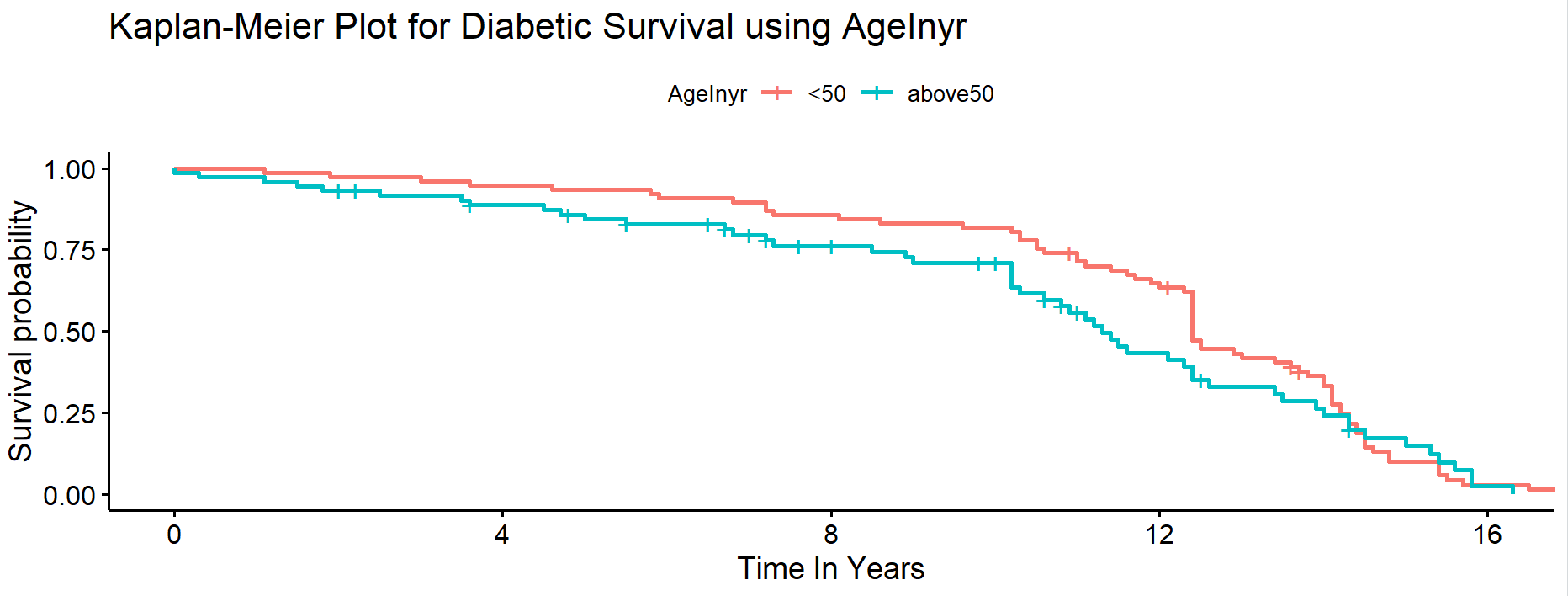
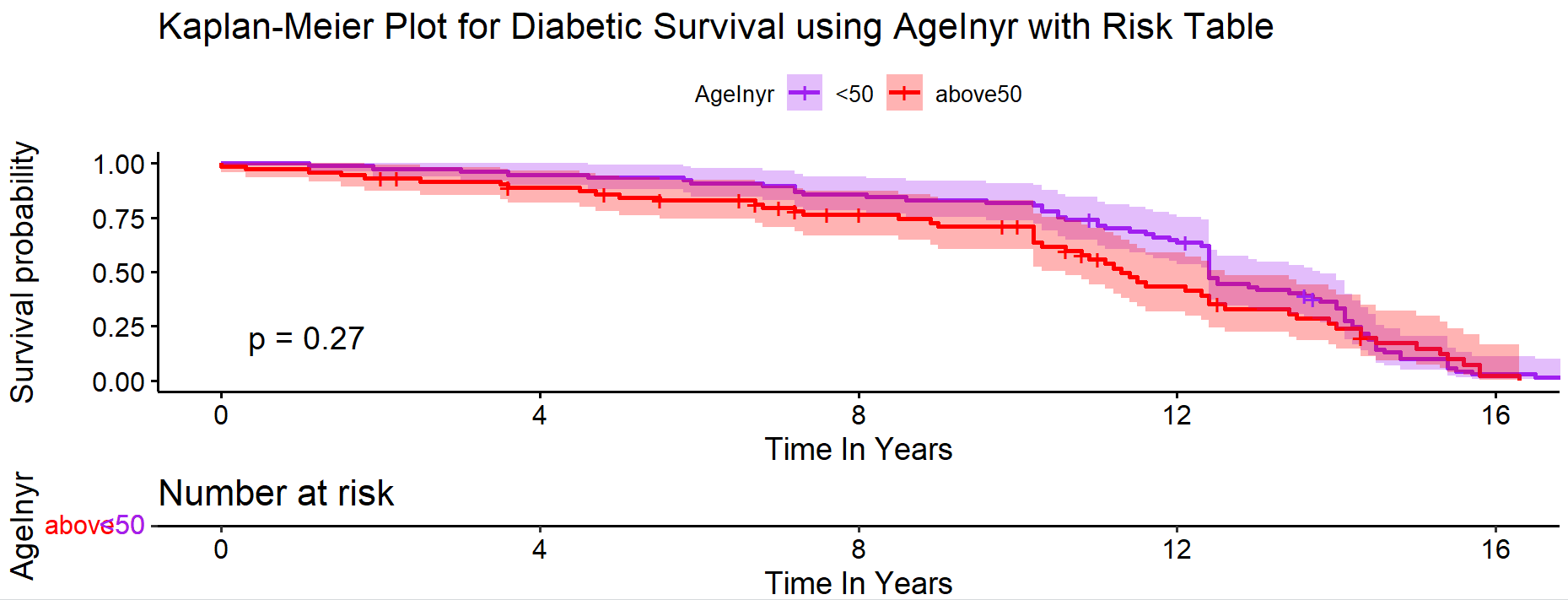


…upto 16.9 years



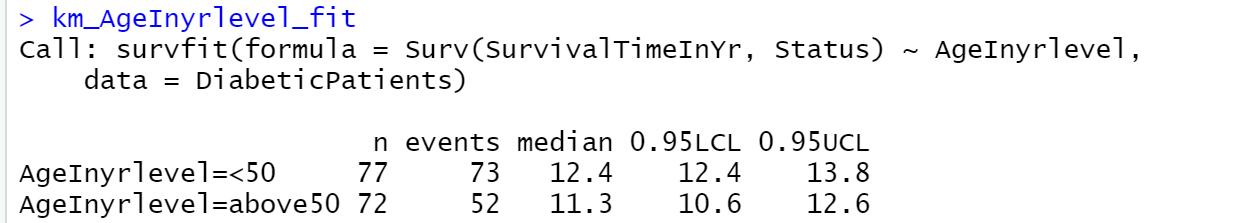
…upto 16.9 years

**7.2.20: Graph Of The Survival Distribution Curve Of Age (In Yrs) At Each Level**

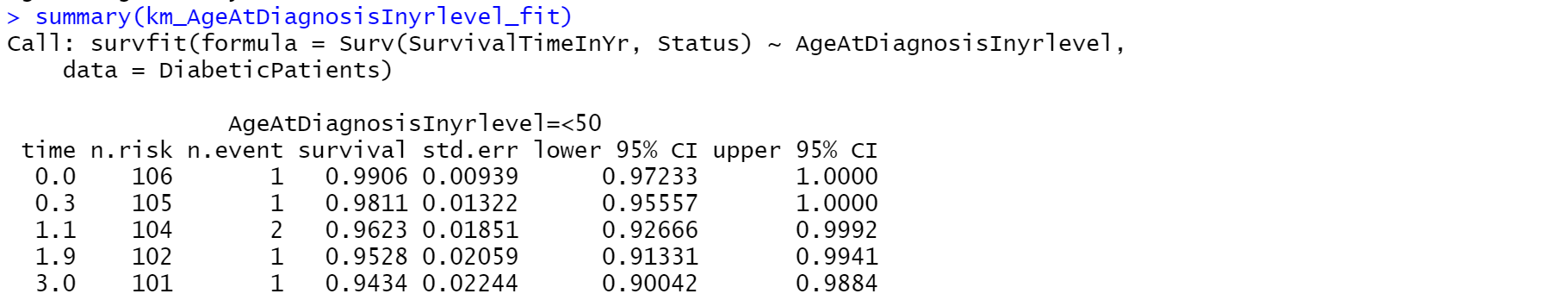
**Comment:**The above are plots of age using two levels. The levels are <50 and above50. It can be seen from the distribution curve that patients that are less than 50 years have higher survival rate than patients that are greater than 50 years.

**7.2.21: The 95% Confidence Interval For The Median Of The Survival Curves Using Age**

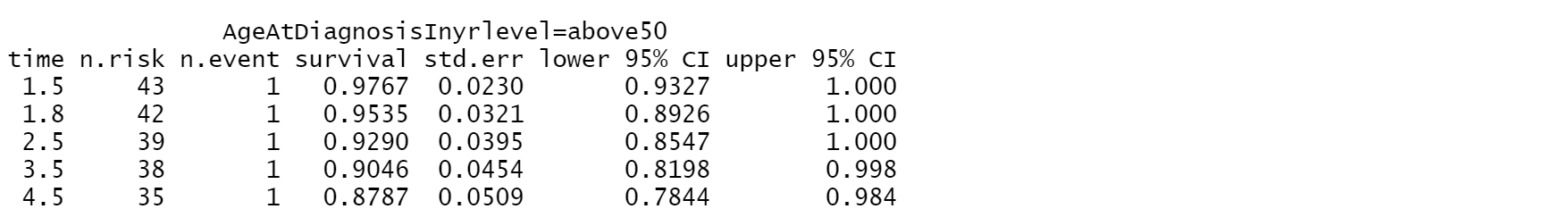


**Comment:**The median survival time for patients of age <50 is 12.4years with a 95% confidence interval of 12.4-13.8 years . Patients of age above 50 is has a median survival time of 11.3 years and a 95% confidence interval of 10.26-12.6 years.

**7.2.22: The Survival Distribution Of Age At Diagnosis(In Yrs) At Each Level**

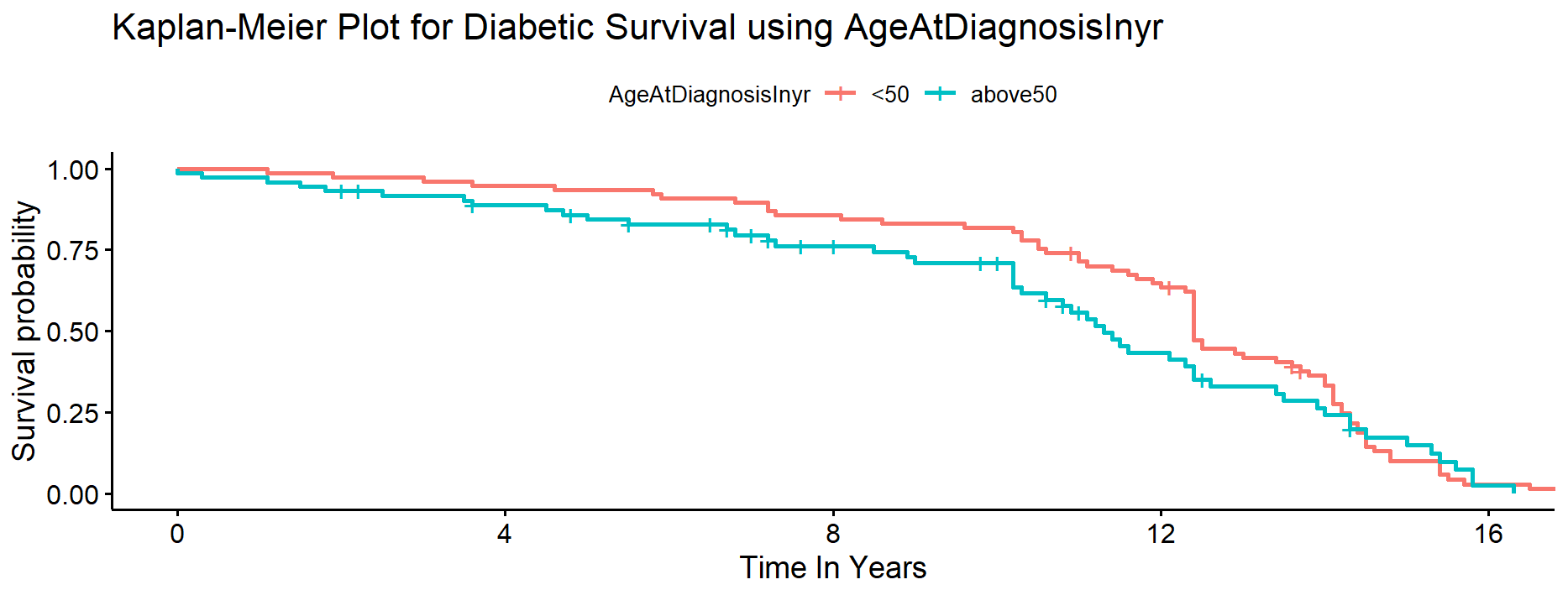


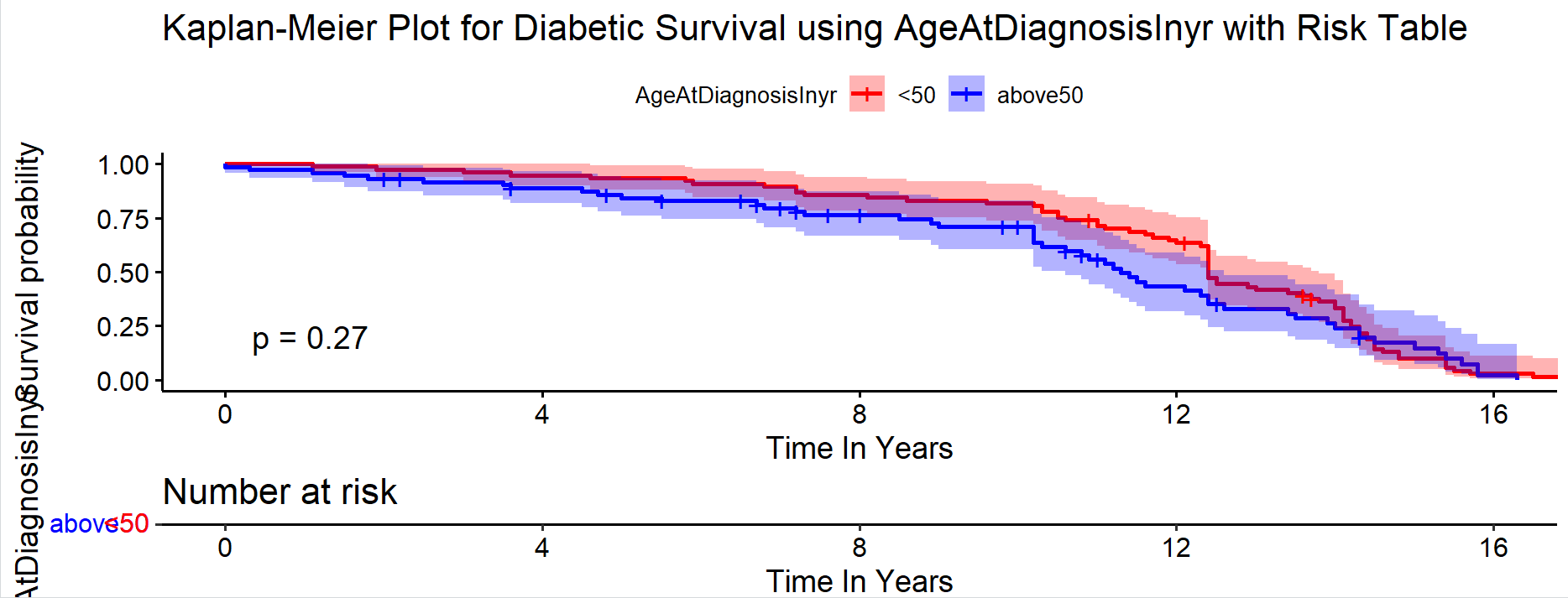
…upto 16.9 years



…upto 16.9 years

**7.2.23: Graph of the Survival Distribution Curve of Age At Diagnosis(In Yrs) at each level**

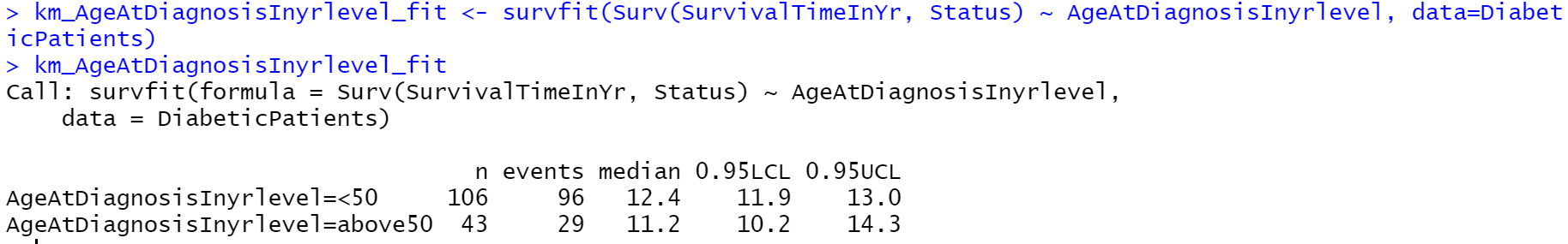




**Comments:**

The above are plots of age using two levels, The levels are <50) and above50. It can be seen from the distribution curve that patients that are less than 50 years at diagnosis have higher survival times than patients that are greater than 50 years.

**7.2.24: The 95% Confidence Interval For The Median Of The Survival Curves Using All Age At Diagnosis**



**Comment:**The median survival time for patients of age at diagnosis <50 is 12.4years with a 95% confidence interval of 11.9- 13 years . Patients of age at diagnosis above 50 has a median survival time of 11.2 years and a 95% confidence interval of 10.2-14.3 years. It can be seen from the distribution curve that patients that are less than 50 years at diagnosis have higher survival times than patients that are greater than 50 years.

**7.3: The Test Of Differences Between The Survival Curves Using The Log Rank Test**

Log Rank test was chosen because we are dealing with weighted variables. It compares the entire survival experience between groups. It can be said to be a test used to know whether the survival curves are identical (overlapping) or not.

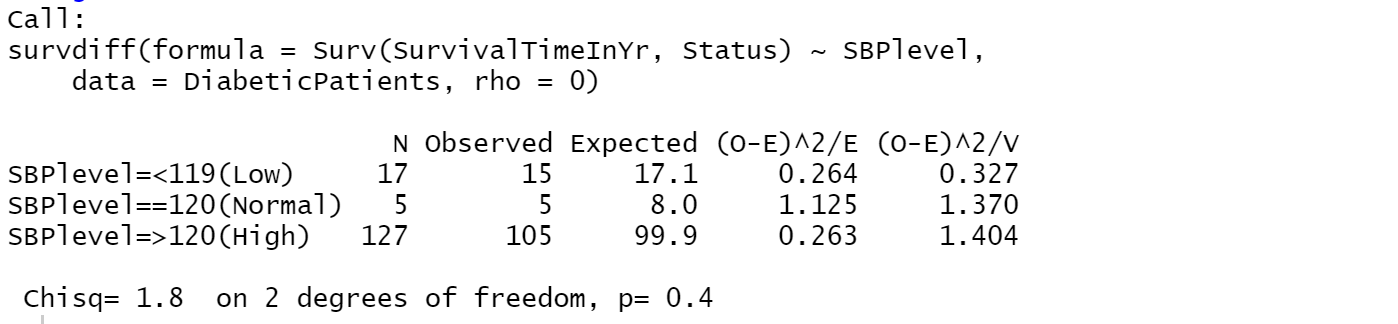
**7.3.1: SBP Group**

**Hypothesis:**

*H0: the survival curves for the three SBP groups are the same*

*Vs*

*Ha: at least one of the survival curves for the three SBP groups is different*



**Comment:** The null hypothesis that there is no difference in the survival curves of the 3 SBP groups against the alternative that at least one of the survival curves for the smoking status groups is different at an alpha level of 0.05. We can see from above that the p-value is 0.4 >0.05 and the chi-square value is 1.8. Therefore, we fail to reject the null hypothesis and conclude that the survival curves for the three SBP groups are the same.

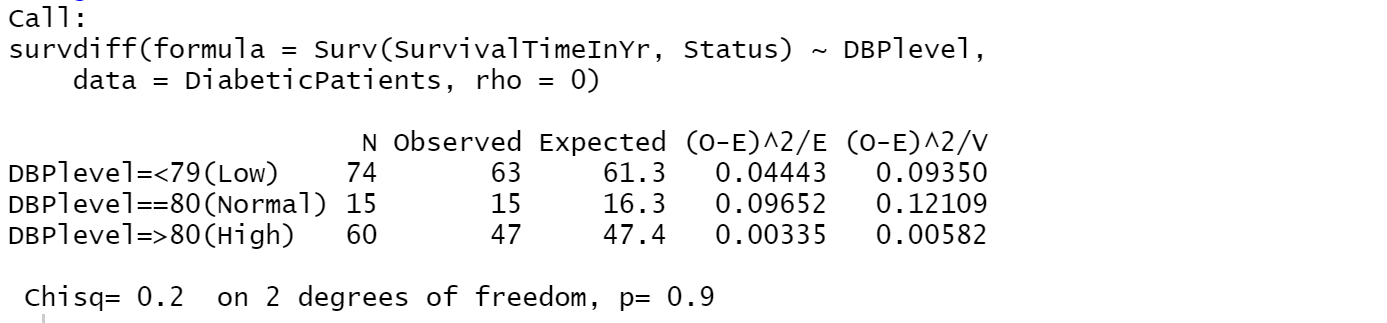
**7.3.2 DBP Group**

**Hypothesis:**

*H0: the survival curves for the three DBP groups are the same*

*Vs*

*Ha: at least one of the survival curves for the three DBP groups is different*



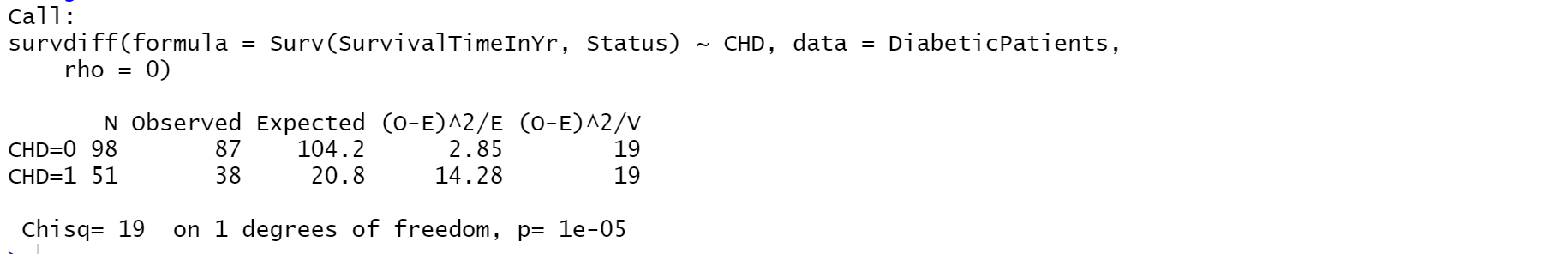
**Comment:** The null hypothesis is that there is no difference in the survival curves of the 3 SBP groups against the alternative that at least one of the survival curves for the smoking status groups is different at an alpha level of 0.05. We can see from above that the p-value is 0.9 >0.05 and the chi-square value is 0.2. Therefore, we fail to reject the null hypothesis and conclude that the survival curves for the three DBP groups are the same.

**7.3.3: CHD group**

**Hypothesis:**

*H0: the survival curves between the two CHD groups are the same*

*Ha: the survival curves between the two CHD groups are different*



**Comment:** The null hypothesis is that survival curves of the CHD groups are the same against the alternative that at the survival curves for the CHD groups are not the same at an alpha level of 0.05. We can see from above that the p-value is 0.00001 <0.05 and the chi-square value is 19. Therefore, we reject the null hypothesis and conclude that the survival curves for the two CHD groups 0(No) and 1 (Yes) are different.

**7.3.4: ECG Group**



**Comment:** The null hypothesis is that there is no difference in the survival curves of the ECG groups against the alternative that at least one of the survival curves for the ECG groups is different at an alpha level of 0.05. We can see from above that the p-value is 0.000004 <0.05 and the chi-square value is 24.9, we to reject the null hypothesis and conclude that at least one of the survival curves for the three ECG status groups are different from each other.

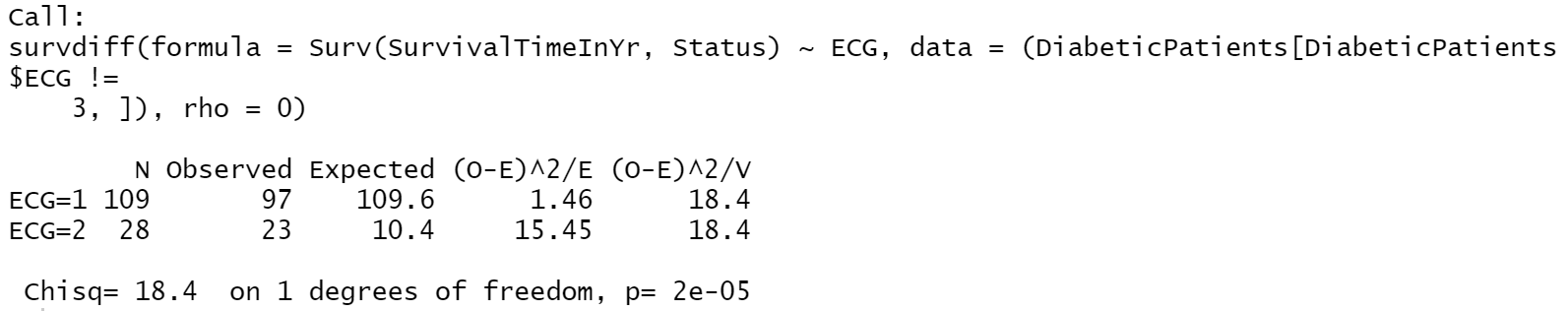
**7.4: Pairwise Tests To Determine If There Is Any Differences In Survival Curves Between Pairs Of Groups**

From the above analysis we rejected the null hypothesis and conclude that at least one of the survival curves for the three ECG status groups are different. Now we are going to use the pairwise test to determine those ECG groups that there survival curves are different from each other.

**7.4.1: Between ECG groups 1 and 2**

**Hypothesis:**

*H0 : the survival curve between the two ECG groups 1 and 2 are the same*

*Ha : the survival curves between the two ECG groups 1 and 2 are not the same* 

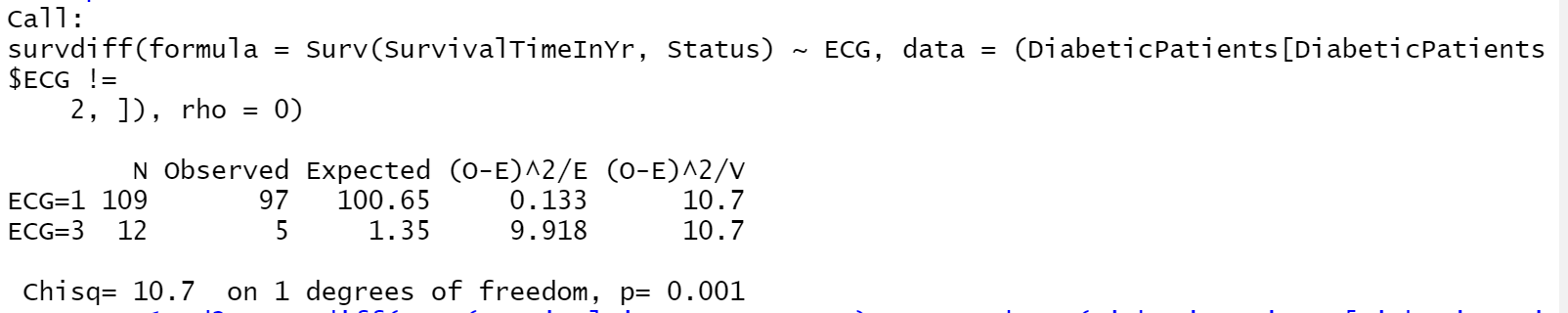
**Comment:** Since the ​ χ22 = 18.4, *p* – *value,* 0.00002 < 0.05, we reject the null hypothesis and conclude that the survival curves between the ECG group 1(Normal) and 2 (Borderline) are not the same.

**7.4.2: Between ECG groups 1 and 3**

**Hypothesis:**

*H0 : the survival curve between the two ECG groups 1 and 3 are the same*

*Ha : the survival curve between the two ECG groups 1 and 3 are not the same*



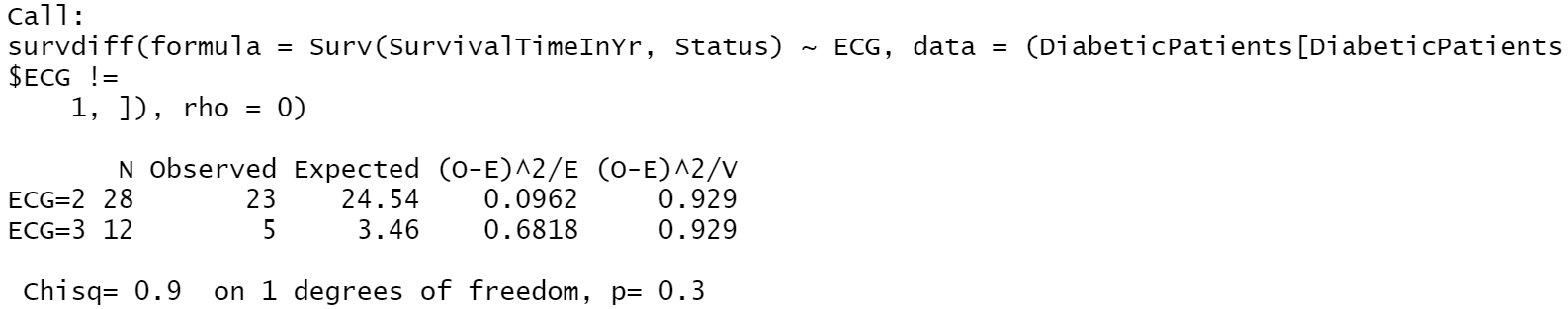
**Comment:** Since the ​ χ22 = 10.7, *p* – *value,* 0.001 < 0.05, we reject the null hypothesis and conclude that the survival curve between groups 1(Normal) and 3 (Abnormal) are not the same.

**7.4.3: Between ECG groups 2 and 3**

**Hypothesis:**

*H0 : the survival curve between the two ECG groups 2 and 3 are the same*

*Ha : the survival curve between the two ECG groups 2 and 3 are not the same*

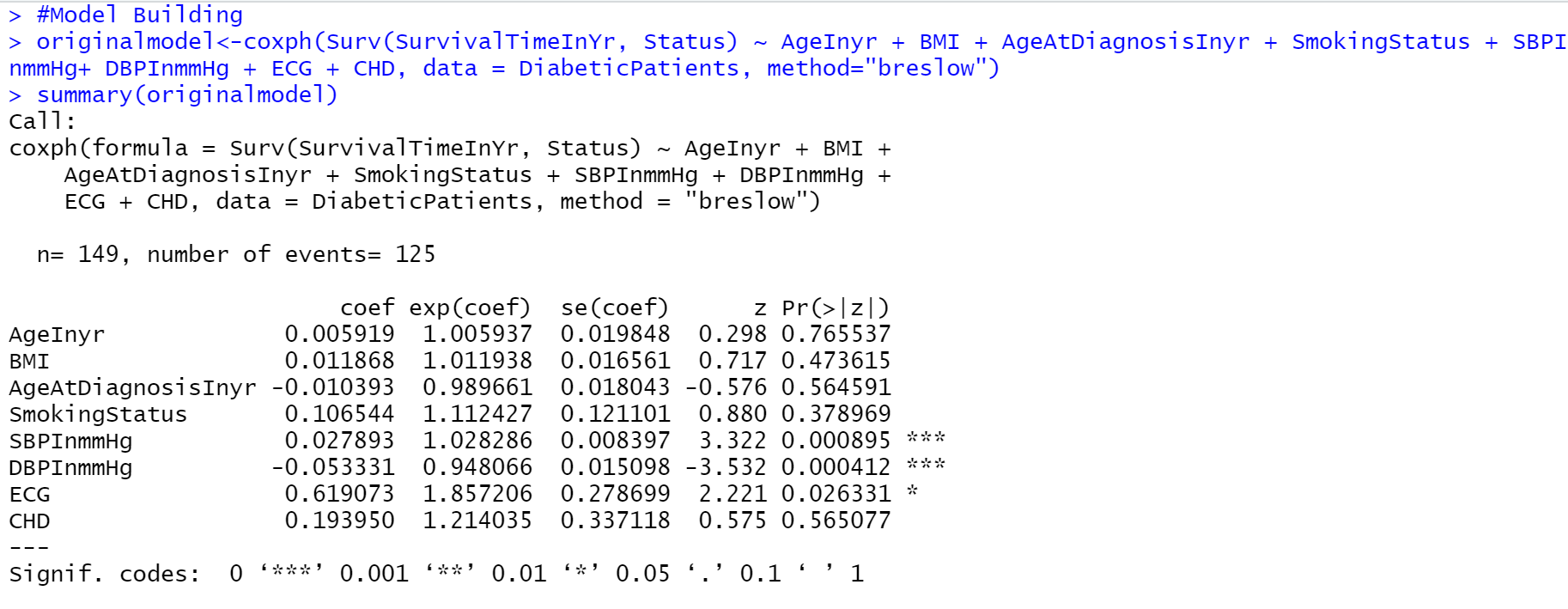


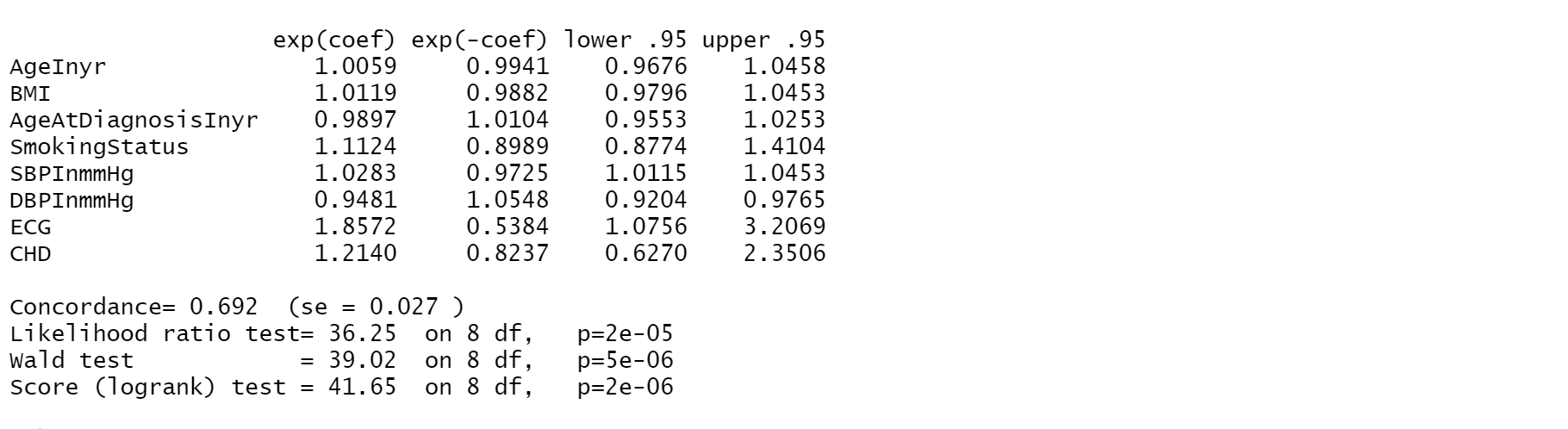
**Comment:** Since the ​ χ22 = 0.9, *p* – *value,* 0.3 > 0.05, we fail to reject the null hypothesis and conclude that the survival between groups 2(Borderline) and 3 (Abnormal) are the same.

**Conclusion:** From the pairwise test we found out that the survival curves pairs of the ECG group 1 and 2 and that of ECG groups 1 and 3 are not the same.

**7.5: The Confidence Interval For The Relative Risk Of Any Pair Of The Categorical Variable Using Cox Proportional Hazard Model**

Now we will use the Cox Proportional Hazard Model to estimate the confidence interval for the relative risk of the pairs of the categorical variable. The choice of Cox Proportional Hazard Model is because of its robustness. Below is the result;



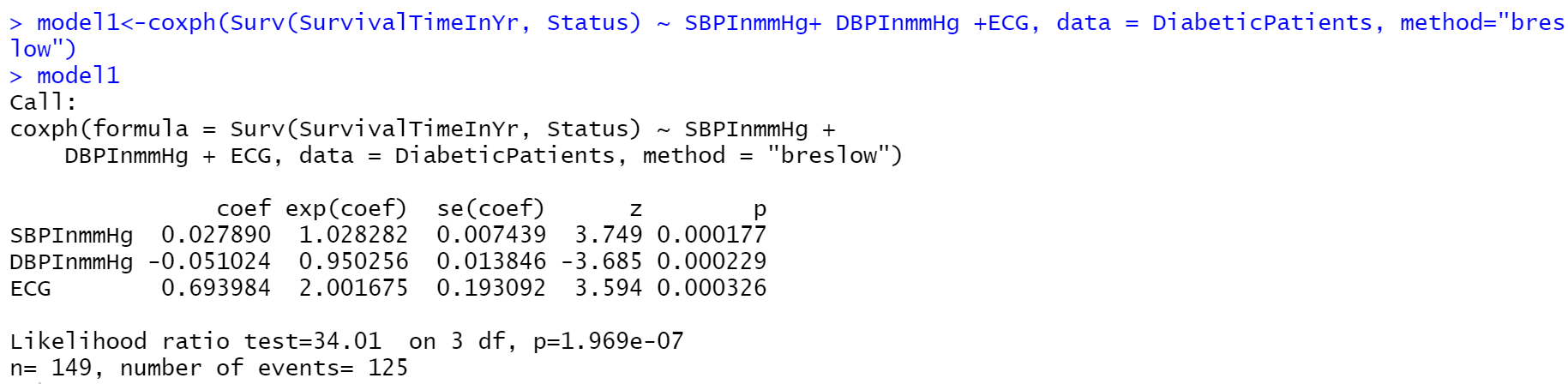


**Comment:** From the Cox Proportional Hazard Model above SBPInmmHG, DBPInmmHg and ECG are significant with a p-values of 0.000895, 0.000412 and 0.026331 respectively. The Likelihood ratio test has a p-value of 0.00002 <0.05 hence, it is significant. The Likelihood Ratio Test was considered because it is more stable than Wald Test and Score (Log Rank) Test.

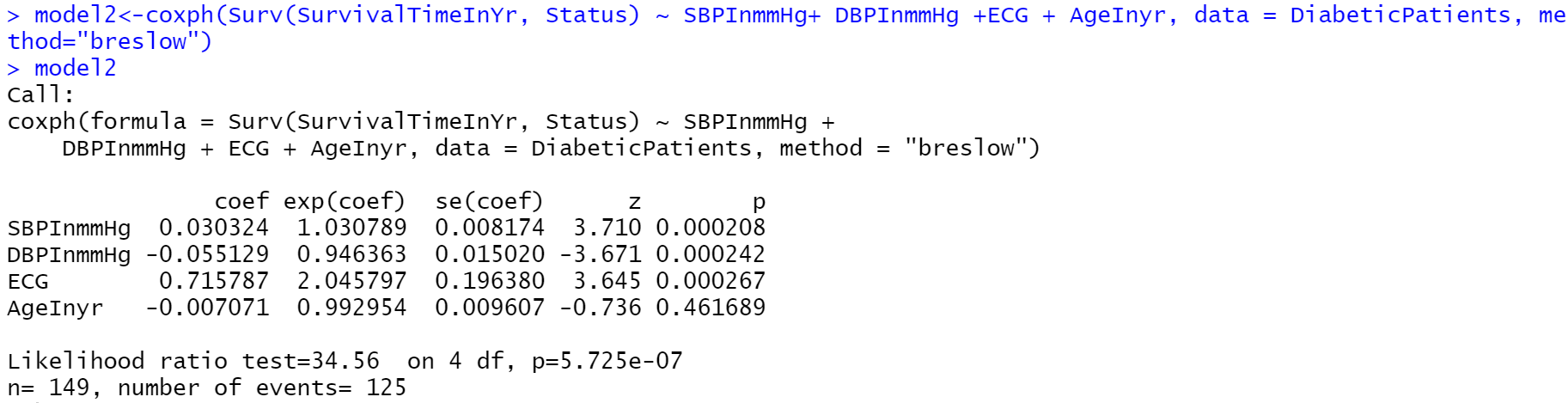
1. **Model Building**

From the Cox Proportional Hazard Model above we have that SBPInmmHG, DBPInmmHg and ECG are significant. Now we are going to start building the model with these 3 significant variables and then do forward selection.

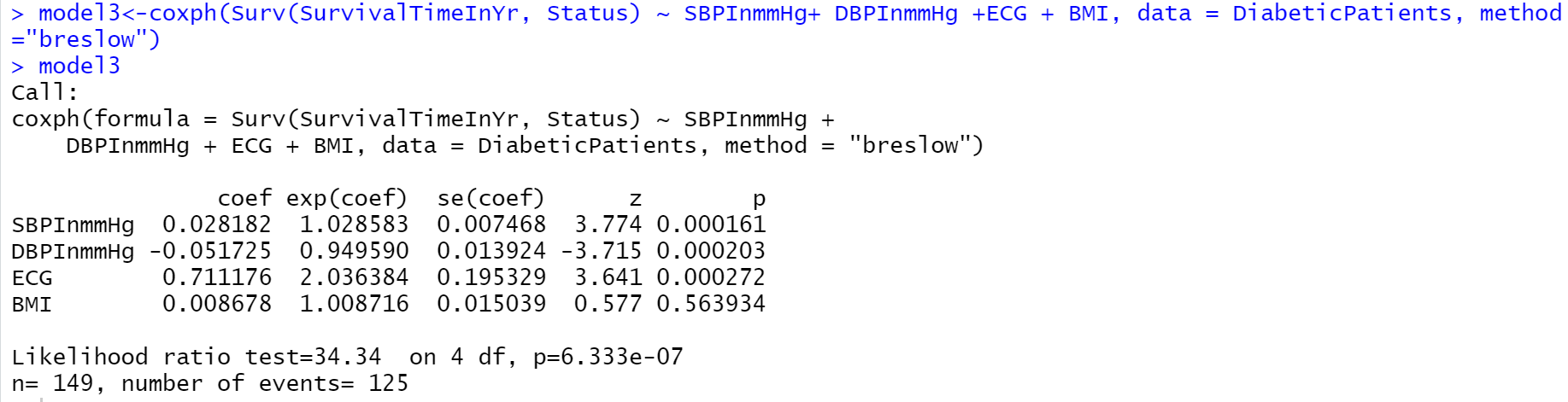
**8.1: Cox Proportional Hazard Model**

**8.1.1: Model with SBPInmmHG, DBPInmmHg and ECG** 

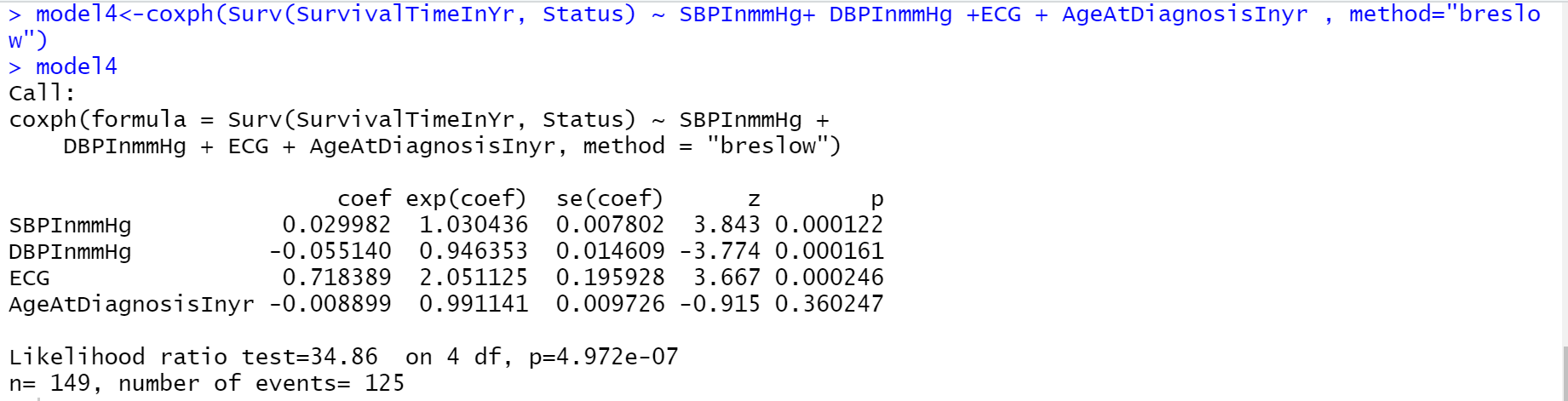
**Comment:** As can be seen above, the model is significant and SBPInmmHG, DBPInmmHg and ECG are also significant individually. Now we add more models to them.

**8.1.2: Model with SBPInmmHG, DBPInmmHg, ECG and AgeInyr**

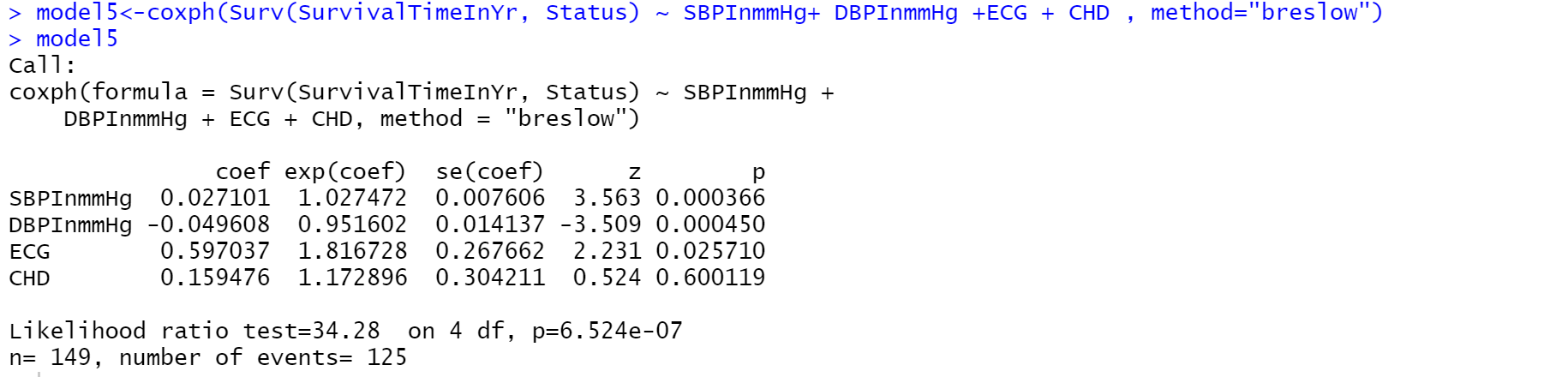
**Comment:** As can be seen above, AgeInyr has no effect on the model. SBPInmmHG, DBPInmmHg and ECG are still significant.

**8.1.3: Model with SBPInmmHG, DBPInmmHg, ECG and BMI**

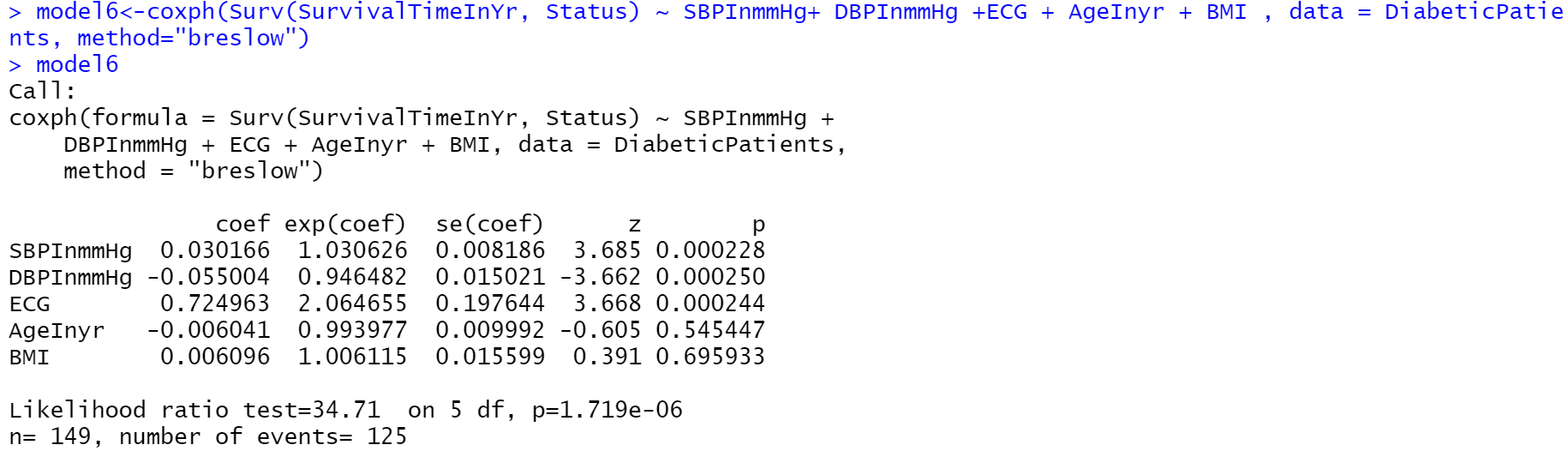
**Comment:** Again, BMI has no effect on the model. AgainSBPInmmHG, DBPInmmHg and ECG are still significant.

**8.1.4: Model with SBPInmmHG, DBPInmmHg, ECG and AgeAtDiagnosis**

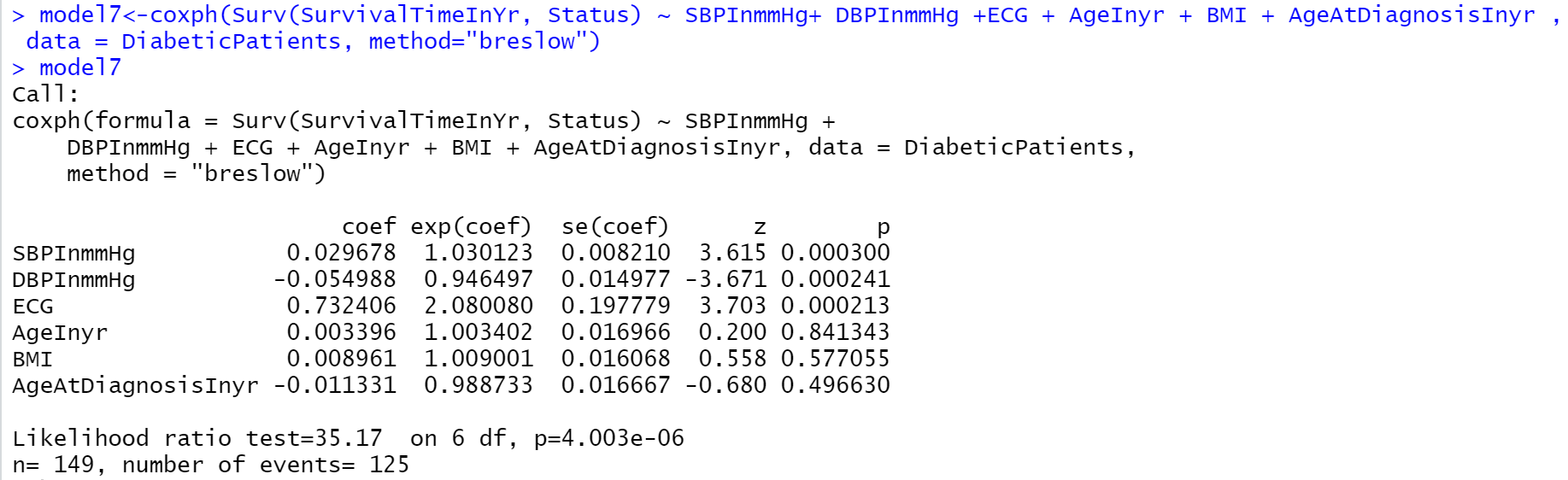
**Comment:** AgeAtDiagnosticInyr has no effect on the model. SBPInmmHG, DBPInmmHg and ECG are significant again.

**8.1.5: Model with SBPInmmHG, DBPInmmHg, ECG and CHD**

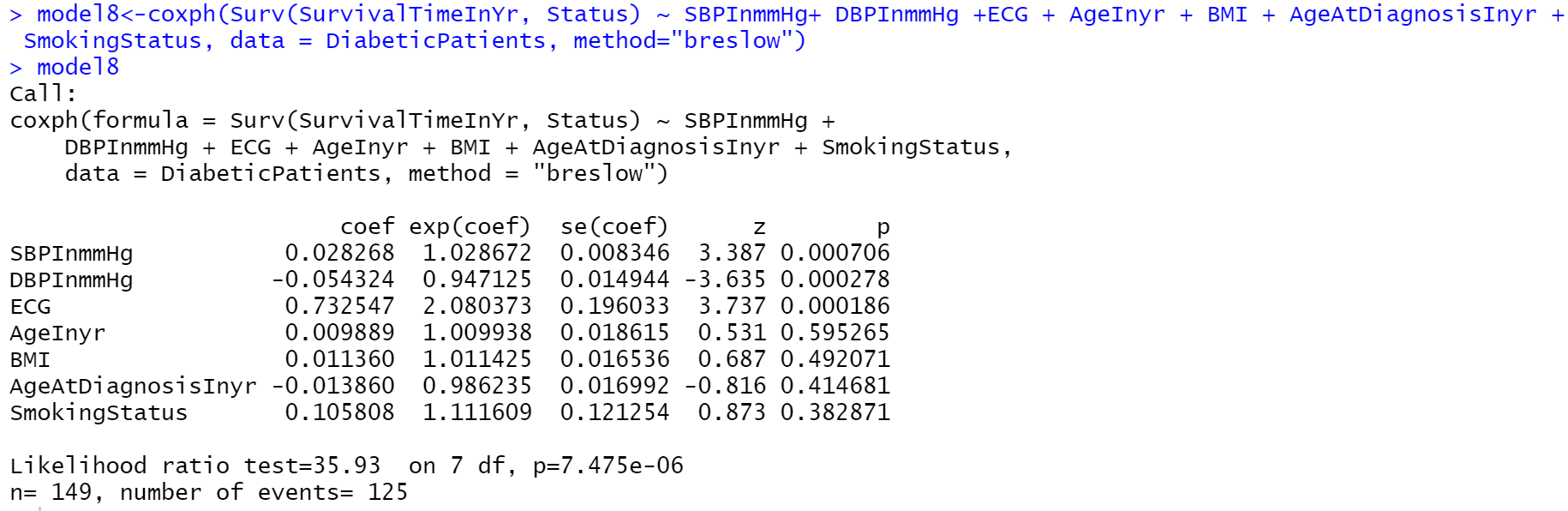
**Comment:** CHD has no effect on the model. SBPInmmHG, DBPInmmHg and ECG are significant.

**8.1.6: Model with SBPInmmHG, DBPInmmHg, ECG, AgeInyr and BMI** 

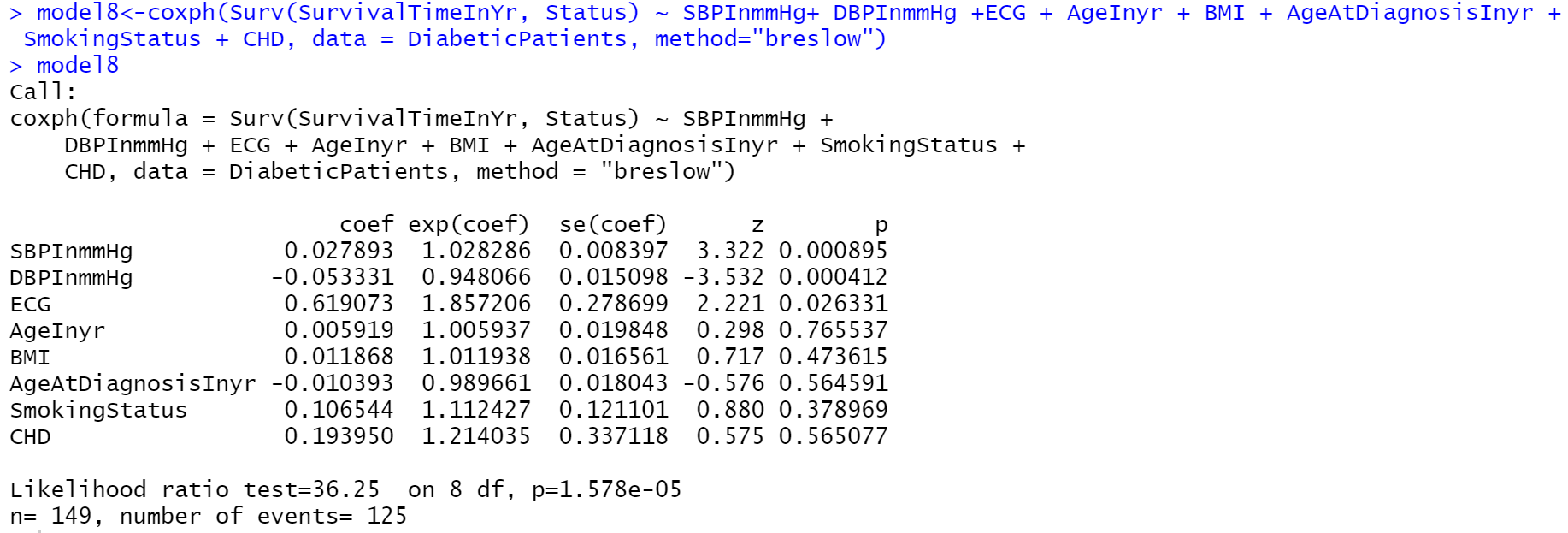
**Comment:** AgeInyr and BMI have no effect on the model. SBPInmmHG, DBPInmmHg and ECG are significant.

**8.1.7: Model with SBPInmmHG, DBPInmmHg, ECG, AgeInyr, BMI and AgeAtDiagnosisInyr**

**Comment:** AgeInyr , BMI and AgeAtDiagnosis have no effect on the model. Only SBPInmmHG, DBPInmmHg and ECG are significant.

**8.1.8: Model with SBPInmmHG, DBPInmmHg, ECG, AgeInyr, BMI and AgeAtDiagnosisInyr and SMokingStatus**

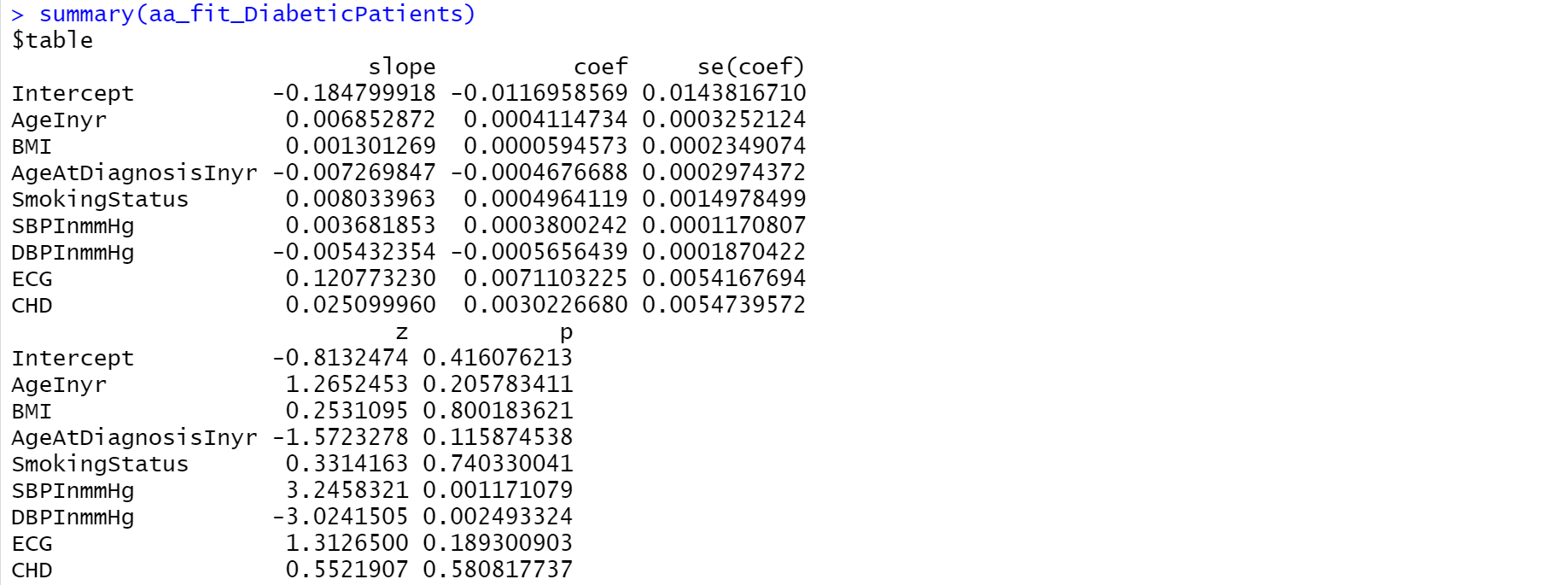
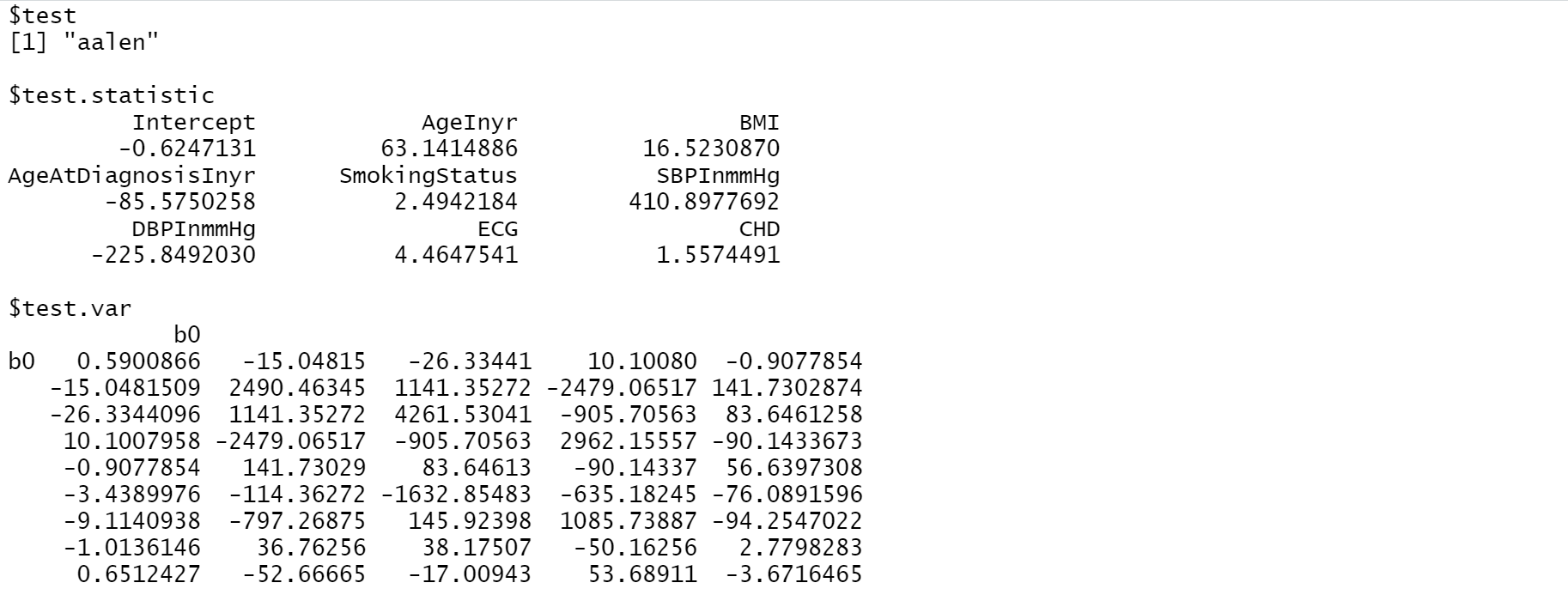
**Comment:** AgeInyr , BMI, AgeAtDiagnosis and SmokingStatus have no effect on the model. Only SBPInmmHG, DBPInmmHg and ECG are significant again.

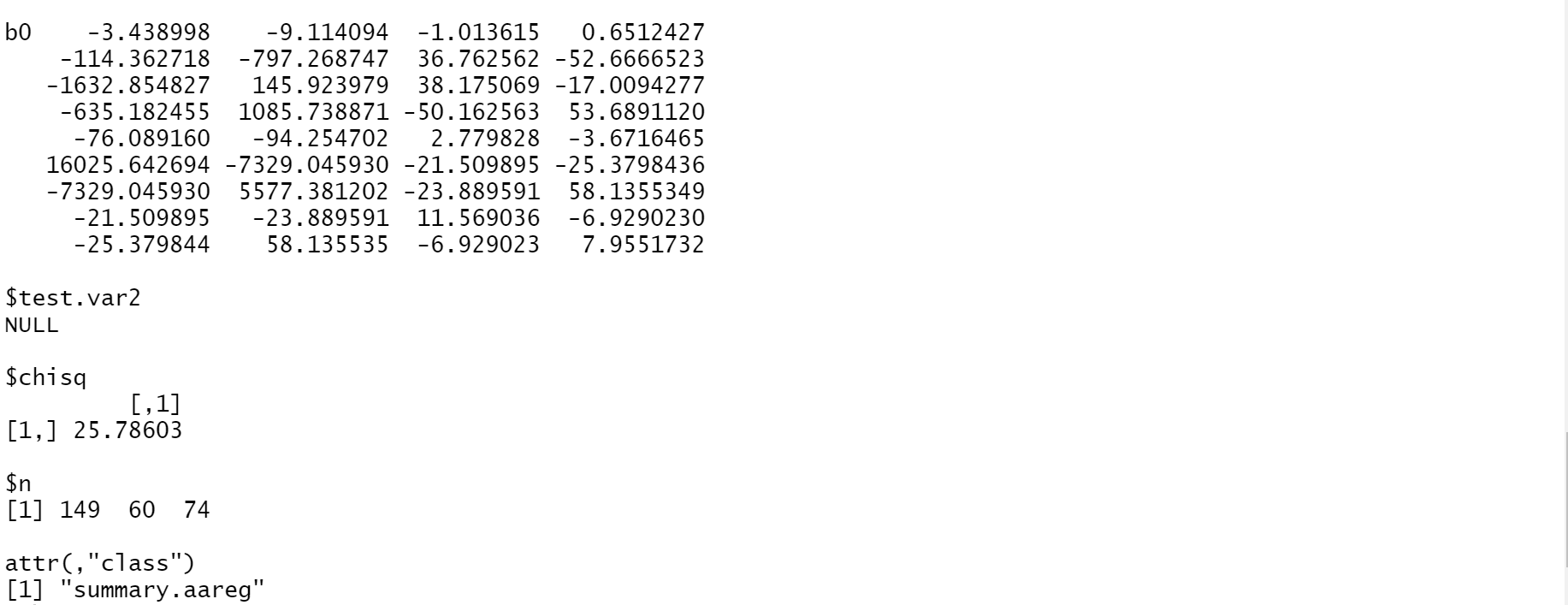
**8.1.9: Model with SBPInmmHG, DBPInmmHg, ECG, AgeInyr, BMI and AgeAtDiagnosisInyr, SMokingStatus** and CHD

**Comment:** AgeInyr , BMI, AgeAtDiagnosis , SmokingStatus and CHD have no effect on the model. Only SBPInmmHG, DBPInmmHg and ECG are significant.

**8.2: Aalen’s Additive Regression Model**

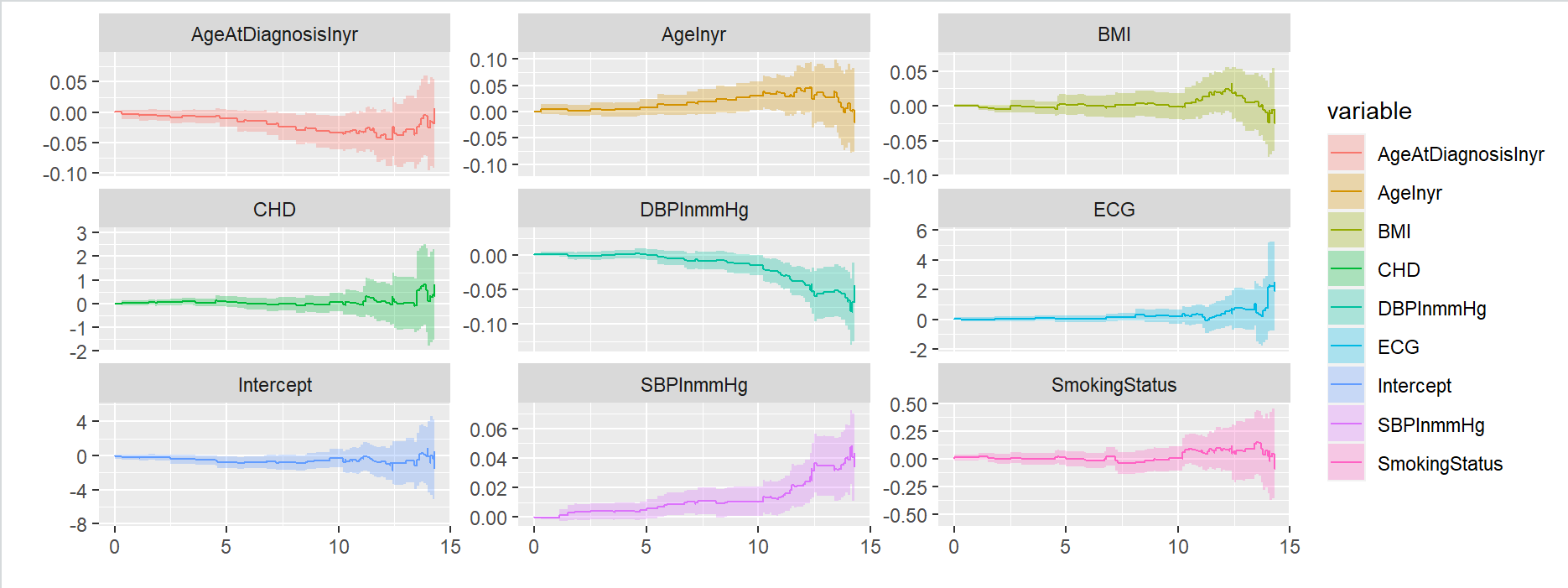
Aalen’s Additive regression Model was also used to fit the diabetic patient dataset. The Aalen model assumes that the cumulative hazard H(t) for a subject can be expressed as a(t) + X B(t), where a(t) is a time-dependent intercept term, X is the vector of covariates for the subject (possibly time-dependent), and B(t) is a time-dependent matrix of coefficients.” and below is the result.

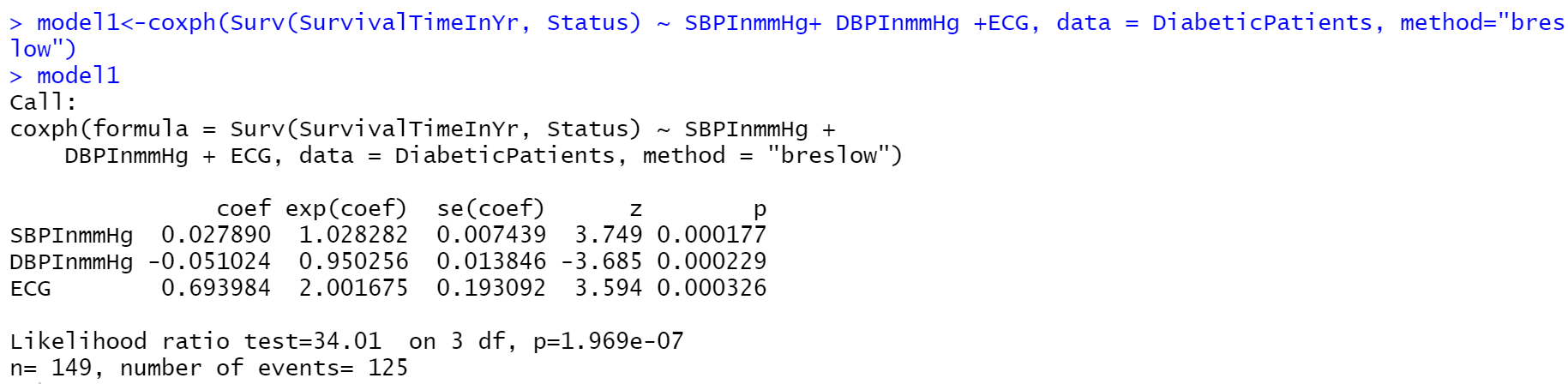


**Comment:** From the Aalen’s Additive Regression Model above and unlike the Cox Proportional Model only SBPInmmHg, DBPInmmHg are significant with a p-values of 0.001171979 and 0.002493324 respectively.

The plots show how the effects of the covariates change over time.



**8.3: Final Model**

Using the cox proportional Hazard model, the final model is determined and it is the model with SBPInmmHG, DBPInmmHg and ECG 

**Comment:**

* After using the forward, backward and stepwise selection the final model is the one with SBPInmmHG, DBPInmmHg and ECG because at every stage of the model they are significant.
* At a given instant in time the probability of dying for patients who have higher Systolic Blood Pressure is almost 3% higher than a patients who have lower Systolic Blood while adjusting for the other variables.
* In the same vein, at a given instant in time, patients who have higher Diastolic Blood Pressure is 5% more likely to survive than patients with lower Diastolic Blood Pressure. It means that at a given instant in time, patient who have high DBP is 0.05% more likely to survive than patients with low DBP.
* At a given instant in time, patient who have Abnormal ElectroCardiogram reading (ECG) is 1.877 times more likely to die than patients with Boarder-line and Normal ECG. It means that at a given instant in time, patients who have Abnormal ECG is 88% more likely to die than patients with Boarder-line and Normal ECG reading.

1. **FINDINGS AND CONCLUSION**

From estimating the survival curves of the variables, to using an appropriate estimator to estimate the confidence interval for the median of their survival curves; to test for difference between their survival curves; to building an appropriate model for the dataset and finally obtaining an estimator and confidence interval for the relative risks one thing stood out and that is - Cox Proportional Hazard Regression Model shows Systolic Blood Pressure (mmHg), Diastolic Blood Pressure (mmHg) and Electrocardiogram (ECG) are significant. That is to say that they have effect on the survival probability and survival time of the 149 patients with diabetics more than the other variables in the dataset.

While Cox Proportional Hazard Regression Model shows that only SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), ECG (Electrocardigram) have significant effect on survival time, Aalen’s Additive Regression Model shows that only SBP (Systolic Blood Pressure) and DBP (Diastolic Blood Pressure) have significant effect on survival time. The former was chosen as the final model because of the robustness of Cox Proportional Hazard Regression Model.

One striking revelation that contradicts popular belief is that smoking doesn’t have a significant effect on survival probability and survival time of the diabetic patients. The BMI doesn’t affect survival time of the diabetic patients and so is age and age at diagnosis. Also it was discovered that whether a patient has Coronary Heart Disease or not doesn’t have any effect on survival time of diabetic patients.

# **References**

Lee, E. T., & Wang, J. W. (2003). *Statistical Method for Survival Analysis.* New Jersey: John Wiley & Sons INC.

Moeschberger, M. .., & Klein, J. .. (2003). *Survival Analysis Techniques for Censored and Truncated Data.* New York: Springer-Verlag New York, Inc.

<https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/>

<https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>

<https://github.com/taiyun/corrplot/issues/10>

<https://bioconnector.github.io/workshops/r-survival.html>

<https://thomaselove.github.io/432-notes/cox-regression-models-for-survival-data-example-2.html>

library(readxl)

library(survival)

library(ranger)

library(ggplot2)

library(dplyr)

library(ggfortify)

library(survminer)

DiabeticPatients<- read\_excel("C:/Users/ogwur/Documents/Spring Courses/Survival Analysis Course Files/Survival Analysis Project File/Survival Analysis Project Main Data.xlsx")

DiabeticPatients

head(DiabeticPatients)

xkm <- with(DiabeticPatients, Surv(SurvivalTimeInYr, Status))

xkm

str(DiabeticPatients)

summary(DiabeticPatients)

?cut

attach(DiabeticPatients)

names(DiabeticPatients)

class(AgeInyr)

AgeInyrlevel<-cut(AgeInyr, breaks=c(0, 50, 90), labels = c("<50", "above50"))

AgeInyrlevel[1:149]

km\_AgeInyrlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ AgeInyrlevel, data=DiabeticPatients)

km\_AgeInyrlevel\_fit

summary(km\_AgeInyrlevel\_fit)

ggsurvplot(km\_AgeInyrlevel\_fit, legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeInyr",title="Kaplan-Meier Plot for Diabetic Survival using AgeInyr")

ggsurvplot(km\_AgeInyrlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeInyr", palette=c("purple", "red"),

title="Kaplan-Meier Plot for Diabetic Survival using Smoking AgeInyr with Risk Table", risk.table.height=.20)

AgeAtDiagnosisInyrlevel<-cut(AgeAtDiagnosisInyr, breaks=c(0, 50, 90), labels = c("<50", "above50"))

AgeAtDiagnosisInyrlevel[1:149]

km\_AgeAtDiagnosisInyrlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ AgeInyrlevel, data=DiabeticPatients)

km\_AgeAtDiagnosisInyrlevel\_fit

summary(km\_AgeAtDiagnosisInyrlevel\_fit)

ggsurvplot(km\_AgeAtDiagnosisInyrlevel\_fit, legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeAtDiagnosisInyr",title="Kaplan-Meier Plot for Diabetic Survival using AgeAtDiagnosisInyr")

ggsurvplot(km\_AgeAtDiagnosisInyrlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeAtDiagnosisInyr", palette=c("Red", "blue"),

title="Kaplan-Meier Plot for Diabetic Survival using AgeAtDiagnosisInyr with Risk Table", risk.table.height=.20)

SBPlevel<-cut(SBPInmmHg, breaks=c(0, 119, 120, 250), labels = c("<119(Low)", "=120(Normal)", ">120(High)"))

SBPlevel[1:149]

km\_SBPlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SBPlevel, data=DiabeticPatients)

km\_SBPlevel\_fit

summary(km\_SBPlevel\_fit)

ggsurvplot(km\_SBPlevel\_fit, legend.labs=c("<119(Low)", "=120(Normal)", ">120(High)"), xlab="Time In Years", legend.title="SBPmmHg",title="Kaplan-Meier Plot for Diabetic Survival using SBP in mmHg")

ggsurvplot(km\_SBPlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<119(Low)", "=120(Normal)", ">120(High)", xlab="Time In Years", legend.title="SBPmmHg", palette=c("Red", "blue", "yellow"),

title="Kaplan-Meier Plot for Diabetic Survival using SBP in mmHg with Risk Table", risk.table.height=.20)

DBPlevel<-cut(DBPInmmHg, breaks=c(0, 79, 80, 120), labels = c("<79(Low)", "=80(Normal)", ">80(High)"))

DBPlevel[1:149]

km\_DBPlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ DBPlevel, data=DiabeticPatients)

km\_DBPlevel\_fit

summary(km\_DBPlevel\_fit)

ggsurvplot(km\_DBPlevel\_fit, legend.labs=c("<79(Low)", "=80(Normal)", ">80(High)"), xlab="Time In Years", legend.title="DBPmmHg",title="Kaplan-Meier Plot for Diabetic Survival using DBP in mmHg")

ggsurvplot(km\_DBPlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<79(Low)", "=80(Normal)", ">80(High)", xlab="Time In Years", legend.title="DBPmmHg", palette=c("Red", "blue", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using DBP in mmHg with Risk Table", risk.table.height=.20)

km\_SBPInmmHg\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg, data=DiabeticPatients)

km\_SBPInmmHg\_fit

summary(km\_SBPInmmHg\_fit)

ggsurvplot(km\_SBPInmmHg\_fit, xlab="Time In Years", legend.title="SDPInmmHg",title="Kaplan-Meier Plot for Diabetic Survival using SBP")

ggsurvplot(km\_SBPInmmHg\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE, xlab="Time In Years", legend.title="SDPInmmHg",

title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status Group with Risk Table", risk.table.height=.20)

km<- survfit(Surv(SurvivalTimeInYr, Status) ~ 1, data=DiabeticPatients)

km

ggsurvplot(km,title="Overall Kaplan-Meier Plot for Diabetic Survival", xlab = "Time In Years")

ggsurvplot(km, conf.int=TRUE, pval=TRUE, xlab = "Time In Years", risk.table=TRUE, palette=c("dodgerblue2"),

title="Overall Kaplan-Meier Plot for Diabetic Survival with Risk Table", risk.table.height=.20)

km\_SmokingStatus\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data=DiabeticPatients)

km\_SmokingStatus\_fit

summary(km\_SmokingStatus\_fit)

ggsurvplot(km\_SmokingStatus\_fit, xlab="Time In Years", legend.title="Smoking Status",title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status")

ggsurvplot(km\_SmokingStatus\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("0", "1", "2"), xlab="Time In Years", legend.title="Smoking Status", palette=c("dodgerblue2", "orchid2", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status Group with Risk Table", risk.table.height=.20)

km\_ECG\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ ECG, data=DiabeticPatients)

km\_ECG\_fit

summary(km\_ECG\_fit)

ggsurvplot(km\_ECG\_fit,legend.title="ECG", xlab="Time In Years", title="Kaplan-Meier Plot for Diabetic Survival using ECG Group")

ggsurvplot(km\_ECG\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("1", "2", "3"), xlab="Time In Years", legend.title="ECG", palette=c("dodgerblue2", "orchid2", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using ECG Group with Risk Table", risk.table.height=.20)

km\_CHD\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ CHD, data=DiabeticPatients)

km\_CHD\_fit

summary(km\_CHD\_fit)

ggsurvplot(km\_CHD\_fit, xlab="Time In Years", legend.title="CHD",title="Kaplan-Meier Plot for Diabetic Survival using CHD Group")

ggsurvplot(km\_CHD\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("0", "1"), xlab="Time In Years", legend.title="CHD", palette=c("dodgerblue2", "orchid2"),

title="Kaplan-Meier Plot for Diabetic Survival using CHD Group with Risk Table", risk.table.height=.20)

logranktestSBPlevel<-survdiff(Surv(SurvivalTimeInYr, Status) ~ SBPlevel, data=DiabeticPatients, rho=0)

logranktestSBPlevel

logranktestDBPlevel<-survdiff(Surv(SurvivalTimeInYr, Status) ~ DBPlevel, data=DiabeticPatients, rho=0)

logranktestDBPlevel

#Comparing the 3 levels of Smoking Status

logranktestSmokingStatus<-survdiff(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data=DiabeticPatients, rho=0)

logranktestSmokingStatus

#Comparing the 3 levels of ECG

logranktestECG<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=DiabeticPatients, rho=0)

logranktestECG

#testing pair difference

compare1and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=3,]), rho=0)

compare1and2

compare0and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=2,]), rho=0)

compare0and2

compare1and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=1,]), rho=0)

compare1and2

#Comparing the 2 levels of CHD

logranktestCHD<-survdiff(Surv(SurvivalTimeInYr, Status) ~ CHD, data=DiabeticPatients, rho=0)

logranktestCHD

coxSmokingStatus<- coxph(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data = DiabeticPatients)

coxSmokingStatus

coxECG<- coxph(Surv(SurvivalTimeInYr, Status) ~ ECG, data = DiabeticPatients)

coxECG

coxCHD<- coxph(Surv(SurvivalTimeInYr, Status) ~ CHD, data = DiabeticPatients)

coxCHD

coxDiabeticPatients<- coxph(Surv(SurvivalTimeInYr, Status) ~ AgeInyr+ BMI + AgeAtDiagnosisInyr + SmokingStatus +

SBPInmmHg + DBPInmmHg + ECG + CHD , data = DiabeticPatients)

summary(coxDiabeticPatients)

coxDiabeticPatients

coxDiabeticPatientsfit <-survfit(coxDiabeticPatients)

coxDiabeticPatientsfit

aa\_fit\_DiabeticPatients <-aareg(Surv(SurvivalTimeInYr, Status) ~ AgeInyr+ BMI + AgeAtDiagnosisInyr + SmokingStatus +

SBPInmmHg + DBPInmmHg + ECG + CHD , data = DiabeticPatients)

aa\_fit\_DiabeticPatients

summary(aa\_fit\_DiabeticPatients)

autoplot(aa\_fit\_DiabeticPatients)

ranger\_DiabeticPatients\_fit <- ranger(Surv(SurvivalTimeInYr, Status) ~ AgeInyr+ BMI + AgeAtDiagnosisInyr + SmokingStatus +

SBPInmmHg + DBPInmmHg + ECG + CHD , data = DiabeticPatients,

mtry = 4,

importance = "permutation",

splitrule = "extratrees",

verbose = TRUE)

ranger\_DiabeticPatients\_fit

death\_times <- ranger\_DiabeticPatients\_fit$unique.death.times

surv\_prob <- data.frame(ranger\_DiabeticPatients\_fit$survival)

avg\_prob <- sapply(surv\_prob,mean)

plot(ranger\_DiabeticPatients\_fit$unique.death.times,ranger\_DiabeticPatients\_fit$survival[1,],

type = "l",

ylim = c(0,1),

col = "red",

xlab = "Years",

ylab = "survival",

main = "Patient Survival Curves")

cols <- colors()

for (n in sample(c(2:dim(vet)[1]), 20)){

lines(ranger\_DiabeticPatients\_fit$unique.death.times, ranger\_DiabeticPatients\_fit$survival[n,], type = "l", col = cols[n])

lines(death\_times, avg\_prob, lwd = 2)

legend(500, 0.7, legend = c('Average = black'))

vi\_DiabeticPatients <- data.frame(sort(round(ranger\_DiabeticPatients\_fit$variable.importance, 4), decreasing = TRUE))

names(vi\_DiabeticPatients) <- "importance"

head(vi\_DiabeticPatients)

vi\_DiabeticPatients

kmi <- rep("KM",length(km$SurvivalTimeInYr))

km\_df <- data.frame(km$SurvivalTimeInYr,km$surv,kmi)

names(km\_df) <- c("Time","Surv","Model")

coxi <- rep("Cox",length(coxDiabeticPatients$SurvivalTimeInYr))

cox\_df <- data.frame(coxDiabeticPatientsfit$SurvivalTimeInYr,coxDiabeticPatientsfit$surv,coxi)

names(cox\_df) <- c("Time","Surv","Model")

rfi <- rep("RF",length(ranger\_DiabeticPatients\_fit$unique.death.times))

rf\_df <- data.frame(ranger\_DiabeticPatients\_fit$unique.death.times,avg\_prob,rfi)

names(rf\_df) <- c("Time","Surv","Model")

plot\_df <- rbind(km\_df,cox\_df,rf\_df)

p <- ggplot(plot\_df, aes(x = Time, y = Surv, color = Model))

p + geom\_line()

fit<-coxph(Surv(SurvivalTimeInYr, Status)~ SBPInmmHg + DBPInmmHg + ECG, data = DiabeticPatients)

fit

library(readxl)

library(survival)

library(ranger)

library(ggplot2)

library(dplyr)

library(ggfortify)

library(survminer)

library(tidyverse)

DiabeticPatients<- read\_excel("C:/Users/ogwur/Documents/Spring Courses/Survival Analysis Course Files/Survival Analysis Project File/Survival Analysis Project Main Data.xlsx")

DiabeticPatients

head(DiabeticPatients)

xkm <- with(DiabeticPatients, Surv(SurvivalTimeInYr, Status))

xkm

str(DiabeticPatients)

summary(DiabeticPatients)

my\_cor\_data <- DiabeticPatients[, c(3,4,5,6,8,9)]

my\_cor\_data

my\_cor\_plot <- cor(my\_cor\_data)

my\_cor\_plot

corrplot(my\_cor\_plot, type = "upper", tl.pos = "td",

method = "circle", tl.cex = 0.5, tl.col = 'black',

order = "hclust", diag = FALSE)

corrplot

attach(DiabeticPatients)

names(DiabeticPatients)

class(AgeInyr)

AgeInyrlevel<-cut(AgeInyr, breaks=c(0, 50, 90), labels = c("<50", "above50"))

AgeInyrlevel[1:149]

km0 <- survfit(Surv(SurvivalTimeInYr, Status) ~ 0)

km0

summary(km0)

km1<- survfit(Surv(SurvivalTimeInYr, Status) ~ 1)

km1

summary(km1)

km\_AgeInyrlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ AgeInyrlevel, data=DiabeticPatients)

km\_AgeInyrlevel\_fit

summary(km\_AgeInyrlevel\_fit)

ggsurvplot(km\_AgeInyrlevel\_fit, legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeInyr",title="Kaplan-Meier Plot for Diabetic Survival using AgeInyr")

ggsurvplot(km\_AgeInyrlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeInyr", palette=c("purple", "red"),

title="Kaplan-Meier Plot for Diabetic Survival using Smoking AgeInyr with Risk Table", risk.table.height=.20)

AgeAtDiagnosisInyrlevel<-cut(AgeAtDiagnosisInyr, breaks=c(0, 50, 90), labels = c("<50", "above50"))

AgeAtDiagnosisInyrlevel[1:149]

km\_AgeAtDiagnosisInyrlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ AgeAtDiagnosisInyrlevel, data=DiabeticPatients)

km\_AgeAtDiagnosisInyrlevel\_fit

summary(km\_AgeAtDiagnosisInyrlevel\_fit)

ggsurvplot(km\_AgeAtDiagnosisInyrlevel\_fit, legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeAtDiagnosisInyr",title="Kaplan-Meier Plot for Diabetic Survival using AgeAtDiagnosisInyr")

ggsurvplot(km\_AgeAtDiagnosisInyrlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<50", "above50"), xlab="Time In Years", legend.title="AgeAtDiagnosisInyr", palette=c("Red", "blue"),

title="Kaplan-Meier Plot for Diabetic Survival using AgeAtDiagnosisInyr with Risk Table", risk.table.height=.20)

SBPlevel<-cut(SBPInmmHg, breaks=c(0, 119, 120, 250), labels = c("<119(Low)", "=120(Normal)", ">120(High)"))

SBPlevel[1:149]

km\_SBPlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SBPlevel, data=DiabeticPatients)

km\_SBPlevel\_fit

summary(km\_SBPlevel\_fit)

ggsurvplot(km\_SBPlevel\_fit, legend.labs=c("<119(Low)", "=120(Normal)", ">120(High)"), xlab="Time In Years", legend.title="SBPmmHg",title="Kaplan-Meier Plot for Diabetic Survival using SBP in mmHg")

ggsurvplot(km\_SBPlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<119(Low)", "=120(Normal)", ">120(High)", xlab="Time In Years", legend.title="SBPmmHg", palette=c("Red", "blue", "yellow"),

title="Kaplan-Meier Plot for Diabetic Survival using SBP in mmHg with Risk Table", risk.table.height=.20)

DBPlevel<-cut(DBPInmmHg, breaks=c(0, 79, 80, 120), labels = c("<79(Low)", "=80(Normal)", ">80(High)"))

DBPlevel[1:149]

km\_DBPlevel\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ DBPlevel, data=DiabeticPatients)

km\_DBPlevel\_fit

summary(km\_DBPlevel\_fit)

ggsurvplot(km\_DBPlevel\_fit, legend.labs=c("<79(Low)", "=80(Normal)", ">80(High)"), xlab="Time In Years", legend.title="DBPmmHg",title="Kaplan-Meier Plot for Diabetic Survival using DBP in mmHg")

ggsurvplot(km\_DBPlevel\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("<79(Low)", "=80(Normal)", ">80(High)", xlab="Time In Years", legend.title="DBPmmHg", palette=c("Red", "blue", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using DBP in mmHg with Risk Table", risk.table.height=.20)

km\_SBPInmmHg\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg, data=DiabeticPatients)

km\_SBPInmmHg\_fit

summary(km\_SBPInmmHg\_fit)

ggsurvplot(km\_SBPInmmHg\_fit, xlab="Time In Years", legend.title="SDPInmmHg",title="Kaplan-Meier Plot for Diabetic Survival using SBP")

ggsurvplot(km\_SBPInmmHg\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE, xlab="Time In Years", legend.title="SDPInmmHg",

title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status Group with Risk Table", risk.table.height=.20)

km<- survfit(Surv(SurvivalTimeInYr, Status) ~ 1, data=DiabeticPatients)

km

ggsurvplot(km,title="Overall Kaplan-Meier Plot for Diabetic Survival", xlab = "Time In Years")

ggsurvplot(km, conf.int=TRUE, pval=TRUE, xlab = "Time In Years", risk.table=TRUE, palette=c("dodgerblue2"),

title="Overall Kaplan-Meier Plot for Diabetic Survival with Risk Table", risk.table.height=.20)

km\_SmokingStatus\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data=DiabeticPatients)

km\_SmokingStatus\_fit

summary(km\_SmokingStatus\_fit)

ggsurvplot(km\_SmokingStatus\_fit, xlab="Time In Years", legend.title="Smoking Status",title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status")

ggsurvplot(km\_SmokingStatus\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("0", "1", "2"), xlab="Time In Years", legend.title="Smoking Status", palette=c("dodgerblue2", "orchid2", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using Smoking Status Group with Risk Table", risk.table.height=.20)

km\_ECG\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ ECG, data=DiabeticPatients)

km\_ECG\_fit

summary(km\_ECG\_fit)

ggsurvplot(km\_ECG\_fit,legend.title="ECG", xlab="Time In Years", title="Kaplan-Meier Plot for Diabetic Survival using ECG Group")

ggsurvplot(km\_ECG\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("1", "2", "3"), xlab="Time In Years", legend.title="ECG", palette=c("dodgerblue2", "orchid2", "green"),

title="Kaplan-Meier Plot for Diabetic Survival using ECG Group with Risk Table", risk.table.height=.20)

km\_CHD\_fit <- survfit(Surv(SurvivalTimeInYr, Status) ~ CHD, data=DiabeticPatients)

km\_CHD\_fit

summary(km\_CHD\_fit)

ggsurvplot(km\_CHD\_fit, xlab="Time In Years", legend.title="CHD",title="Kaplan-Meier Plot for Diabetic Survival using CHD Group")

ggsurvplot(km\_CHD\_fit, conf.int=TRUE, pval=TRUE, risk.table=TRUE,

legend.labs=c("0", "1"), xlab="Time In Years", legend.title="CHD", palette=c("dodgerblue2", "orchid2"),

title="Kaplan-Meier Plot for Diabetic Survival using CHD Group with Risk Table", risk.table.height=.20)

logranktestSBPlevel<-survdiff(Surv(SurvivalTimeInYr, Status) ~ SBPlevel, data=DiabeticPatients, rho=0)

logranktestSBPlevel

logranktestDBPlevel<-survdiff(Surv(SurvivalTimeInYr, Status) ~ DBPlevel, data=DiabeticPatients, rho=0)

logranktestDBPlevel

#Comparing the 3 levels of Smoking Status

logranktestSmokingStatus<-survdiff(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data=DiabeticPatients, rho=0)

logranktestSmokingStatus

#Comparing the 3 levels of ECG

logranktestECG<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=DiabeticPatients, rho=0)

logranktestECG

#testing pair difference

compare1and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=3,]), rho=0)

compare1and2

compare0and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=2,]), rho=0)

compare0and2

compare1and2<-survdiff(Surv(SurvivalTimeInYr, Status) ~ ECG, data=(DiabeticPatients[DiabeticPatients$ECG!=1,]), rho=0)

compare1and2

#Comparing the 2 levels of CHD

logranktestCHD<-survdiff(Surv(SurvivalTimeInYr, Status) ~ CHD, data=DiabeticPatients, rho=0)

logranktestCHD

coxSmokingStatus<- coxph(Surv(SurvivalTimeInYr, Status) ~ SmokingStatus, data = DiabeticPatients)

coxSmokingStatus

coxECG<- coxph(Surv(SurvivalTimeInYr, Status) ~ ECG, data = DiabeticPatients)

coxECG

coxCHD<- coxph(Surv(SurvivalTimeInYr, Status) ~ CHD, data = DiabeticPatients)

coxCHD

cox<- coxph(Surv(SurvivalTimeInYr, Status) ~ AgeInyr+ BMI + AgeInyrlevel + AgeAtDiagnosisInyrlevel + SmokingStatus +

SBPlevel + DBPlevel + ECG + CHD , data = DiabeticPatients)

cox

summary(cox)

cox1<- coxph(Surv(SurvivalTimeInYr, Status) ~ AgeInyr+ BMI + AgeInyr + AgeAtDiagnosisInyr + SmokingStatus +

SBPInmmHg + DBPInmmHg + ECG + CHD , data = DiabeticPatients)

cox1

summary(cox1)

cox2<- coxph(Surv(SurvivalTimeInYr, Status) ~ AgeInyrlevel+ BMI + AgeAtDiagnosisInyrlevel + SmokingStatus +

SBPlevel + DBPlevel + ECG + CHD , data = DiabeticPatients)

cox2

summary(cox2)

cox\_fit <-survfit(cox)

cox\_fit

#Model Building

originalmodel<-coxph(Surv(SurvivalTimeInYr, Status) ~ AgeInyr + BMI + AgeAtDiagnosisInyr + SmokingStatus + SBPInmmHg+ DBPInmmHg + ECG + CHD, data = DiabeticPatients, method="breslow")

summary(originalmodel)

model1<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG, data = DiabeticPatients, method="breslow")

model1

model2<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeInyr, data = DiabeticPatients, method="breslow")

model2

model3<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + BMI, data = DiabeticPatients, method="breslow")

model3

model4<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeAtDiagnosisInyr , method="breslow")

model4

model5<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + CHD , method="breslow")

model5

model6<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeInyr + BMI , data = DiabeticPatients, method="breslow")

model6

model7<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeInyr + BMI + AgeAtDiagnosisInyr , data = DiabeticPatients, method="breslow")

model7

model8<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeInyr + BMI + AgeAtDiagnosisInyr + SmokingStatus, data = DiabeticPatients, method="breslow")

model8

model8<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG + AgeInyr + BMI + AgeAtDiagnosisInyr + SmokingStatus + CHD, data = DiabeticPatients, method="breslow")

model8

#Final Model

FinalModel<-coxph(Surv(SurvivalTimeInYr, Status) ~ SBPInmmHg+ DBPInmmHg +ECG , data = DiabeticPatients, method="breslow")

FinalModel