An analysis of the Mayo Clinic Primary Biliary Cirrhosis dataset

Survival Analysis with R

by Polycarp Odawo and Olga Mushakarara,

Data Science Tech Institute A18 cohort

Description of the data

The dataset used for this project is the primary biliary cirrhosis (pbc) from the R software survival package. It contains data collected from a group of patients suffering from liver cirrhosis during a mayo clinic trial between the years 1974 and 1984.

A total of 312 patients participated in the trial from which patients were randomly selected to be administered the drug D-penicillamine. Patients who did not receive the trial drug were instead administered a placebo drug.

In addition basic measurements from 112 liver cirrhosis patients were also collected and included in the dataset. These patients did not take part in the trial, therefore they were neither administered with the trial drug nor the placebo. Six of these patients were lost during follow up, hence the dataset does not have any data concerning them.

Basic descriptive Statistics on the dataset

The raw pbc dataset found in the survival package has a total of 418 observations and 20 variables. The variables are described in the table below.

Variable	Data type	Description
age	int	patient age in years
albumin	int	serum albumin (g/dl)
alk.phos	int	alkaline phosphatase (U/liter)
ascites	int	presence of ascites (0 not present, 1 present)
ast	int	aspartate aminotransferase, once called SGOT (U/ml)
bili	int	amount of serum bilirubin (mg/dl)
chol	int	amount of serum cholesterol (mg/dl)
copper	int	amount of copper in urine (ug/day)
edema	int	0 for no edema, 0.5 for untreated or successful treated and 1 for edema despite diuretic therapy
hepeto	int	presence of an enlarged liver (0 not present, 1 present)
id	char	patient identifier (unique to every patient)
platelet	int	platelet count in the blood
protime	int	standardised blood clotting time (in sec)

Variable	Data type	Description
sex	factor	patient gender (male or female)
spiders	int	blood vessel malformations in the skin(0 not present, 1 present)
stage	int	histological stage of liver cirrhosis in the patient
status	int	Status of the patient at endpoint. (0 for censored, 1 for transplant and 2 for dead.)
time	int	time between patient's start of the trial until death in days
trt	int	treatment type (1 for D-penicillmain, 2 for placebo and N/A for not randomised)
trig	int	triglycerides (mg/dl)

The minimum age recorded is 26 while the maximum age recorded is 74. There were more women who took part in the exercise than men, with the sex variable having a proportion of 44 males to 374 females.

The histogram of the survival time did not show any predefined distribution. The shortest survival period was 41 days while the longest survival period was 4795 days.

At the trial's endpoint there were 232 censored patients, 25 patients received a transplant and 161 patients had died.

We observed that from the status variable there were three possible out comes for any given patient. For analysis purposes we decided to choose the death of a patient as our event of interest and introduced a new variable "newStatus" to indicate this. This resulted in a proportion of 161 deaths versus 257 censored patients at the end of the trial.

A second variable that was introduced is the "group" variable. Each patient was classified depending on the type of treatment received during the ten year period. Proportions for this variable are as follows: 158 patients received the D-penicillamain drug, 106 patients were not randomised (i.e received no drug/didn't participate in the trial) and 154 patients received the placebo drug.

Questions asked.

1. What is the mean survival time for all the patients?

As observed with the Kaplan-Meier curve, the probability of survival gradually decreases as time increases. The median survival time for all the three groups of patients is 9.30

years with a 95% confidence interval, with a lower bound and upper bound of 8.45 and 10.51 years respectively.

2. Do the three groups differ from each other?

Comparing the three groups using the logrank test gives a p-value of 0.9. This p-value is above the 5% threshold, therefore the null hypothesis is not rejected and all the three groups are considered too be the same.

The variables sex, stage, edema and hepato are each used to perform a stratified logrank test. The p-values obtained are; 0.9, 0.8, 0.7 and 0.06 respectively. Since they are all above the 5% threshold the null hypothesis is not rejected. However, the stratified test for hepato is 6% and may suggest that the groups are not that much different from each other. This may warrant further analysis.

3. Can we quantify the difference between the three groups?

The cox regression model gives an opportunity to determine if the groups differ from each other and also quantifies by how much they differ. In this case cox regression model gives a score (logrank) test with p-value 0.9 which is within the same as the logrank test p-value. Hence the null hypothesis is not rejected.

The D-penicillamain group is used as a reference in order to quantify any difference with the other two groups. Not-randomised group has an exp(coef) of 1.03 with a 95% confidence interval, 0.69 lower bound and 1.54 upper bound. This implies that the not-randomised group has a 3% higher risk than the D-penicillamain group. Therefore within this group patients have a shorter time to death.
The placebo group has an exp(-coef) of 1.05 with a 95% confidence interval, 0.67 lower bound and 1.35 upper bound. Therefore, the placebo group have a 5% less risk than the D-penicillamain group, patients in this group have a longer time to death than the D-penicillamain group.

4. What is the effect of age on the progression of the disease?

Since age is a continuous covariate we are able to determine and quantify its effect on the time to death by using cox regression. A p-value of 5.92e-07 is obtained for the age covariate. This value is way below the 5% threshold, hence, it has a significant effect to the survival time of a patient with liver cirrhosis. The exp(coef) value is 1.04 with a 95% confidence interval, 1.02 lower bound and 1.06 upper. For a unit increase in age (years) the risk of death increases by 4%.

Conclusion

The drug D-penicillamain did not offer any added benefits to the patients that were randomly chosen and had it administered to them. This conclusion is supported by the fact that both the logrank test and the cox regression had p-values that are insignificant and the median survival times for the three groups are almost the same as seem with the Kaplan Meier curves.

R code

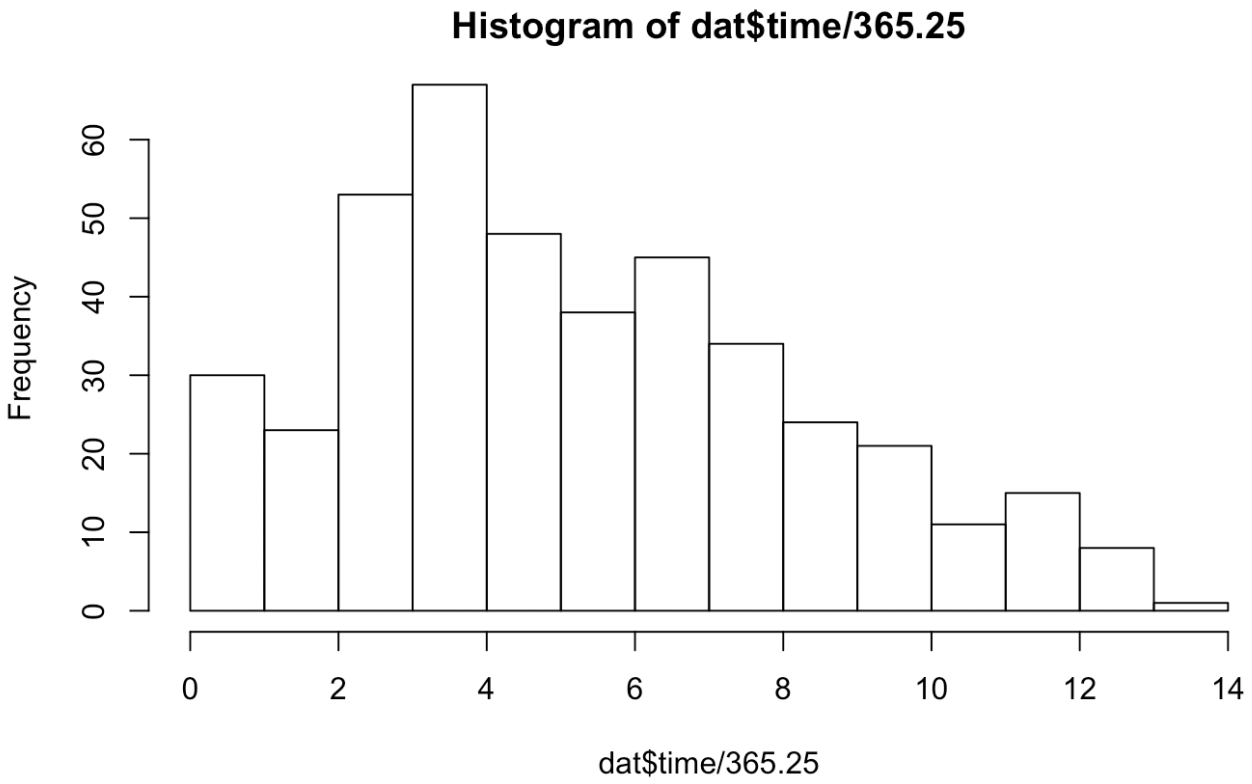
```{r}
library(survival)
library(tidyverse)
dat <- pbc
head(dat)
```

```{r}
summary(dat$time / 365.25)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1123	2.9918	4.7365	5.2506	7.1554	13.1280

```{r}
hist(dat$age)
summary(dat$age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
26.28	42.83	51.00	50.74	58.24	78.44



```{r}
table(dat$sex)
```

```
 m   f
44 374
```

distribution of the time to death in years

```{r}
hist(dat$time/365.25)
summary(dat$time)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
  41    1093   1730 1918    2614 4795
```

```{r}
table(dat$status)
```

```
  0   1   2
232  25 161
```

data preparation and identifying the censoring variable

```{r}
dat$newStatus <- ifelse(dat$status != 2, dat$newStatus <- 0, dat$newStatus <- 1)
dat$trt <- ifelse(!is.na(dat$trt), dat$trt <- dat$trt, dat$trt <- 3)
head(dat$status)
head(dat$newStatus)
dat$group <- ifelse(dat$trt == 1, dat$group <- 'D-penicillmain', ifelse(dat$trt == 2,
dat$group <- 'placebo', dat$group <- 'not-randomised'))
dat$chol = ifelse(is.na(dat$chol),
                  ave(dat$chol, FUN = function(x) mean(x, na.rm = TRUE)),
                  dat$chol)
dat$albumin = ifelse(is.na(dat$albumin),
                  ave(dat$albumin, FUN = function(x) mean(x, na.rm = TRUE)),
                  dat$albumin)
dat$copper = ifelse(is.na(dat$copper),
                  ave(dat$copper, FUN = function(x) mean(x, na.rm = TRUE)),
                  dat$copper)
```

```
dat$alk.phos = ifelse(is.na(dat$alk.phos),
                ave(dat$alk.phos, FUN = function(x) mean(x, na.rm = TRUE)),
                dat$alk.phos)
dat$ast = ifelse(is.na(dat$ast),
                ave(dat$ast, FUN = function(x) mean(x, na.rm = TRUE)),
                dat$ast)
dat$trig = ifelse(is.na(dat$trig),
                ave(dat$trig, FUN = function(x) mean(x, na.rm = TRUE)),
                dat$trig)
dat$platelet = ifelse(is.na(dat$platelet),
                ave(dat$platelet, FUN = function(x) mean(x, na.rm = TRUE)),
                dat$platelet)
dat$protime = ifelse(is.na(dat$protime),
                ave(dat$protime, FUN = function(x) mean(x, na.rm = TRUE)),
                dat$protime)
dat$ascites = ifelse(!is.na(dat$ascites), dat$ascites, dat$ascites <- 0)
dat$hepato = ifelse(!is.na(dat$hepato), dat$hepato, dat$hepato <- 1)
dat$spiders = ifelse(!is.na(dat$spiders), dat$spiders, dat$spiders <- 0)
dat$stage = ifelse(!is.na(dat$stage), dat$stage, dat$stage <- 2)
```

```{r}
table(dat$group)
```

```
D-penicillmain    not-randomised       placebo
         158               106               154
```

proportion of newStatus variable

```{r}
table(dat$newStatus)
```

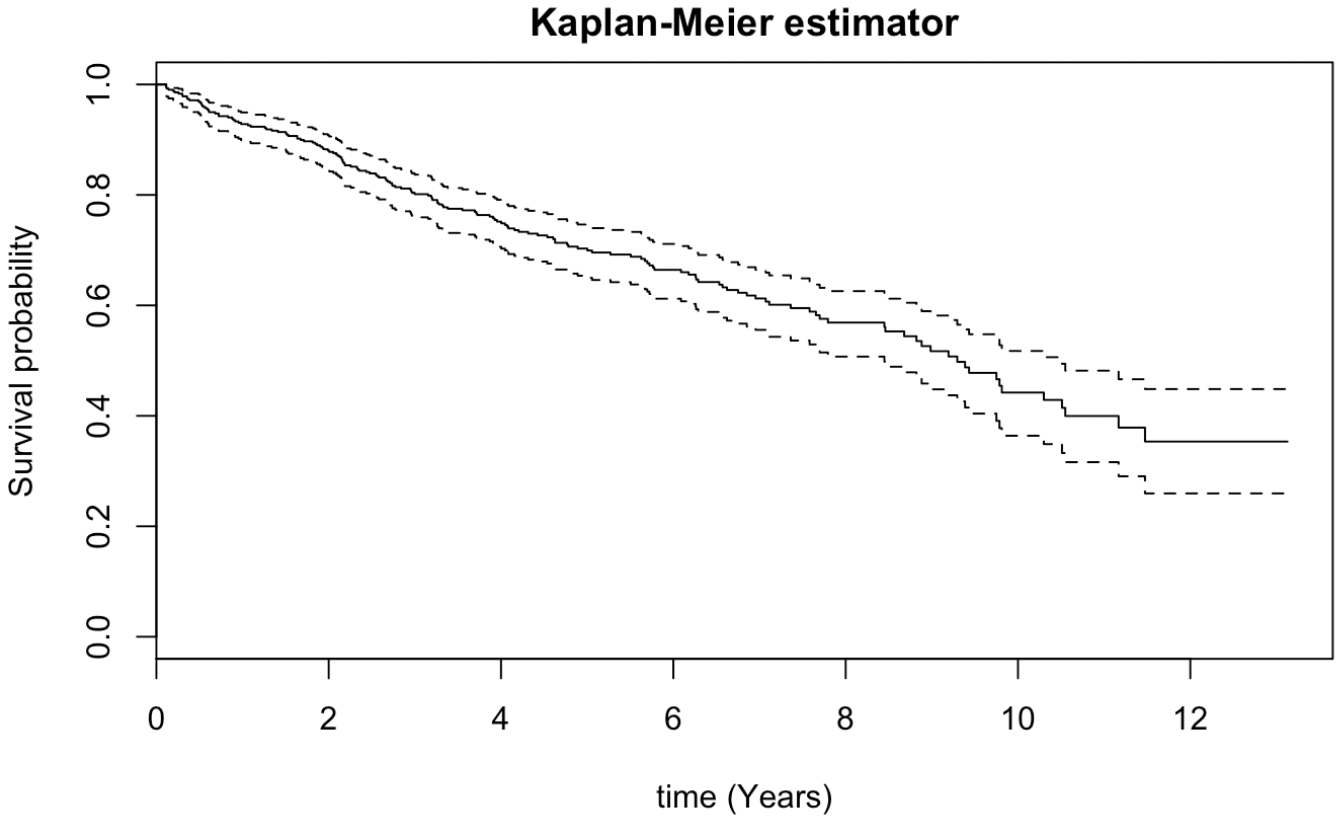
```
  0   1
257 161
```

Kaplan Meyer Estimator

```{r}
dat <- mutate(dat, timeYears = time / 365.25)
fit.KM <- survfit(Surv(timeYears, newStatus) ~ 1, data = dat, conf.type = "log-log").
summary(fit.KM)
plot(fit.KM,
```

```
  main = "Kaplan-Meier estimator",
  ylab = "Survival probability",
  xlab = "time (Years)")
```

```
Call: survfit(formula = Surv(timeYears, newStatus) ~ 1, data = dat,
    conf.type = "log-log")

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.112    418       2    0.995 0.00338        0.981        0.999
 0.118    416       1    0.993 0.00413        0.978        0.998
 0.140    415       1    0.990 0.00476        0.975        0.996
 0.194    414       1    0.988 0.00532        0.972        0.995
 0.211    413       1    0.986 0.00582        0.968        0.994
 0.257    412       1    0.983 0.00628        0.965        0.992
 0.301    411       1    0.981 0.00670        0.962        0.990
 0.304    410       1    0.978 0.00710        0.959        0.989
 0.356    409       1    0.976 0.00747        0.956        0.987
 0.359    408       1    0.974 0.00783        0.953        0.985
 0.383    407       1    0.971 0.00817        0.950        0.984
 0.490    406       1    0.969 0.00849        0.947        0.982
 0.509    405       1    0.967 0.00880        0.944        0.980
 0.523    404       1    0.964 0.00910        0.941        0.978
 0.528    403       1    0.962 0.00938        0.938        0.976
 0.542    402       1    0.959 0.00966        0.935        0.975
 0.567    401       1    0.957 0.00993        0.933        0.973
 0.591    400       1    0.955 0.01019        0.930        0.971
 0.605    399       1    0.952 0.01044        0.927        0.969
 0.611    398       1    0.950 0.01068        0.924        0.967
 0.682    397       1    0.947 0.01092        0.921        0.965
```



what is the median survival time

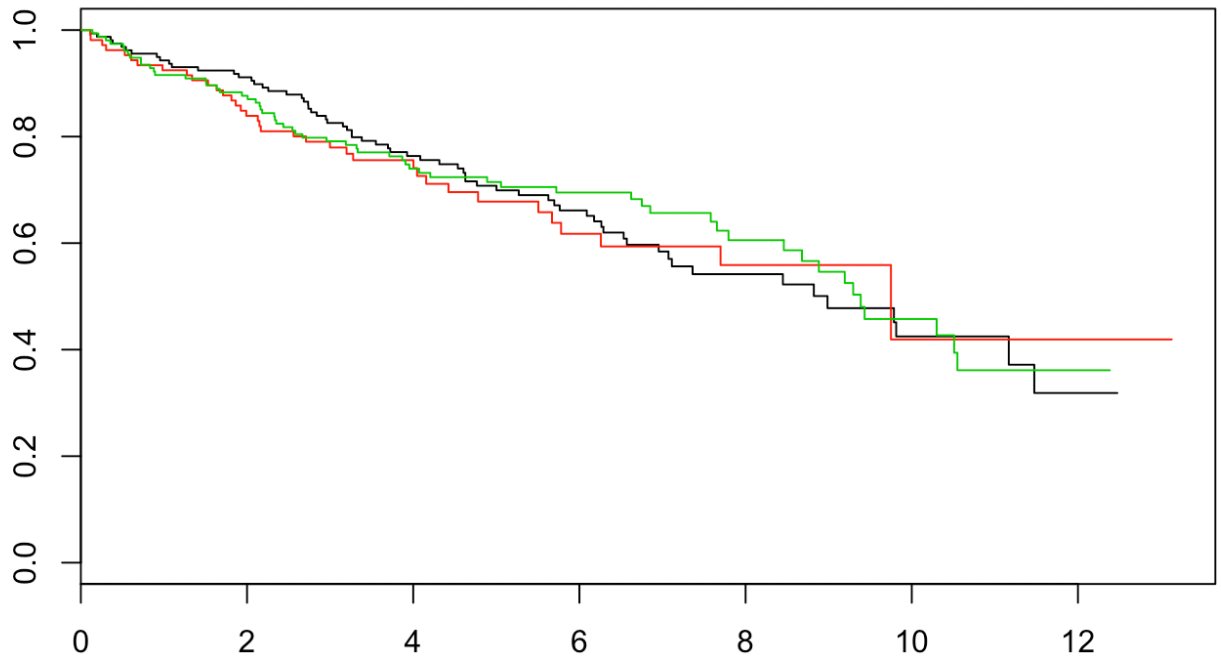
```{r}
fit.KM
```

```
Call: survfit(formula = Surv(timeYears, newStatus) ~ 1, data = dat,
    conf.type = "log-log")

      n  events  median 0.95LCL  0.95UCL
 418.00 161.00    9.30    8.45      10.51
```

comparing three groups D-penicillmain, placebo and not randomised
The logrank test

```{r}
fit.logrank <- survfit(Surv(timeYears, newStatus) ~ group, data = dat, conf.type = "log-log")
fit.logrank
plot(fit.logrank, col = 1:3)
```

```
Call: survfit(formula = Surv(timeYears, newStatus) ~ group, data = dat,
    conf.type = "log-log")

                                n events median 0.95LCL 0.95UCL
group=D-penicillmain          158     65   8.99    6.95    1.1.5
group=not-randomised          106     36   9.75    5.78      NA
```



group=placebo 154 60 9.39 8.46 10.5

Logrank test stratifying by sex

```{r}
survdiff(Surv(time, newStatus) ~ group + strata(sex), data = dat)
```

Call:
survdiff(formula = Surv(time, newStatus) ~ group + strata(sex),
data = dat)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=D-penicillmain	158	65	64.3	0.00869	0.0145
group=not-randomised	106	36	34.5	0.06551	0.0846
group=placebo	154	60	62.3	0.08135	0.1334

Chisq= 0.2 on 2 degrees of freedom, p= 0.9

Logrank test stratifying by holistic stage of disease

```{r}
survdiff(Surv(time, newStatus) ~ group + strata(stage), data = dat)
```

Call:
survdiff(formula = Surv(time, newStatus) ~ group + strata(stage),
data = dat)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=D-penicillmaine	158	65	63.6	0.029	0.0489
group=not-randomised	106	36	33.2	0.239	0.3121
group=placebo	154	60	64.2	0.271	0.4645

Chisq= 0.6 on 2 degrees of freedom, p= 0.8

Logrank test startifying by edema

```{r}
survdiff(Surv(time, newStatus) ~ group + strata(edema), data = dat)
```

Call:
survdiff(formula = Surv(time, newStatus) ~ group + strata(edema),
data = dat)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=D-penicillmain	158	65	65.2	0.000717	0.00123
group=not-randomised	106	36	32.0	0.488430	0.64327
group=placebo	154	60	63.7	0.219444	0.38151

Chisq= 0.8 on 2 degrees of freedom, p= 0.7

Logrank test sytratifying by hepato

```{r}
survdiff(Surv(time, newStatus) ~ group + strata(hepato), data = dat)
```

Call:
survdiff(formula = Surv(time, newStatus) ~ group + strata(hepato),
    data = dat)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=D-penicillmain	158	65	55.5	1.629	2.599
group=not-randomised	106	36	48.2	3.069	5.043
group=placebo	154	60	57.4	0.122	0.195

Chisq= 5.5 on 2 degrees of freedom, p= 0.06

cox regression ==> exp(beta) refers to unit increase of the regressor. exp(beta) is the hazard ratio . from exp(coef) the not-randomised group has a higher risk, therefore shorter life span. placebo has a lower risk hence longer life span. From both p-values it's clear that the 3 sets of groups are not statistically different from each other.

```{r}
fit.cph <- coxph(Surv(time, newStatus) ~ group, data = dat)
summary(fit.cph)
```

Call:
coxph(formula = Surv(time, newStatus) ~ group, data = dat)

  n= 418, number of events= 161

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
groupnot-randomised	0.02560	1.02593	0.20851	0.123	0.902
groupplacebo	-0.05349	0.94792	0.17913	-0.299	0.765

	exp(coef)	exp(-coef)	lower .95	upper .95
groupnot-randomised	1.0259	0.9747	0.6818	1.544
groupplacebo	0.9479	1.0549	0.6673	1.347

Concordance= 0.512 (se = 0.024)
Rsquare= 0   (max possible= 0.985)
Likelihood ratio test= 0.16 on 2 df,   p=0.9
Wald test   = 0.16 on 2 df,   p=0.9
Score (logrank) test = 0.16 on 2 df,   p=0.9

running cox regression on a continuous covariate. From the p-value it is clear that age is a statistically significatnt covariate in how the patients respond to the various treatment methods. The HR is 1.04 which implies that as a patient grows older by 1 year the risk of death increases by 4 percent. the scale of the covariate is given in years. +ve coeffient ==> beta the higher the age the higher the risk of dying

```{r}
fit.cph1 <- coxph(Surv(time, newStatus) ~ age, data = dat)
summary(fit.cph1)
```

Call:
coxph(formula = Surv(time, newStatus) ~ age, data = dat)

  n= 418, number of events= 161

       coef   exp(coef) se(coef)   z   Pr(>|z|)
age 0.039185 1.039963   0.007847   4.994   5.92e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95   upper .95
age   1.04   0.9616   1.024   1.056

Concordance= 0.616 (se = 0.023)
Rsquare= 0.058   (max possible= 0.985)
Likelihood ratio test= 25.19 on 1 df,   p=5e-07
Wald test   = 24.94 on 1 df,   p=6e-07
Score (logrank) test = 25.3 on 1 df,   p=5e-07

Automatic model selection based on AIC

```{r}
Mfull <- coxph(Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + spiders + stage + albumin + group + copper + alk.phos + ast + trig + platelet + bili + chol + edema + protime, data = dat)
```

```{r}
MAIC <- step(Mfull)
```

Start: AIC=1561.21
Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + spiders +

```
    stage + albumin + group + copper + alk.phos + ast + trig +
    platelet + bili + chol + edema + protime

           Df    AIC
- group     2 1557.5
- spiders   1 1559.2
- platelet  1 1559.2
- alk.phos  1 1559.2
- chol      1 1559.7
- sex       1 1559.7
- ascites   1 1560.1
- hepato    1 1560.2
<none>        1561.2
- trig      1 1561.2
- copper    1 1562.9
- ast       1 1563.1
- edema     1 1564.8
- protime   1 1565.0
- albumin   1 1566.6
- age       1 1570.2
- stage     1 1570.9
- bili      1 1585.6

Step:  AIC=1557.49
Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + spiders +
    stage + albumin + copper + alk.phos + ast + trig + platelet +
    bili + chol + edema + protime

           Df    AIC
- spiders   1 1555.5
- platelet  1 1555.5
- alk.phos  1 1555.5
- sex       1 1555.9
- chol      1 1556.0
- ascites   1 1556.3
- hepato    1 1557.0
- trig      1 1557.3
<none>        1557.5
- ast       1 1559.3
- copper    1 1559.3
- edema     1 1561.0
- protime   1 1561.4
- albumin   1 1563.3
- stage     1 1567.1
- age       1 1567.7
- bili      1 1581.6

Step:  AIC=1555.49
Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + stage +
    albumin + copper + alk.phos + ast + trig + platelet + bili +
```

```
   chol + edema + protime

         Df    AIC
- platelet  1 1553.5
- alk.phos  1 1553.5
- sex       1 1553.9
- chol      1 1554.0
- ascites   1 1554.3
- hepato    1 1555.0
- trig      1 1555.3
<none>        1555.5
- ast       1 1557.3
- copper    1 1557.5
- edema     1 1559.0
- protime   1 1559.5
- albumin   1 1561.3
- age       1 1565.8
- stage     1 1566.0
- bili      1 1580.3

Step:  AIC=1553.5
Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + stage +
   albumin + copper + alk.phos + ast + trig + bili + chol +
   edema + protime

         Df    AIC
- alk.phos  1 1551.5
- sex       1 1551.9
- chol      1 1552.0
- ascites   1 1552.3
- hepato    1 1553.0
- trig      1 1553.3
<none>        1553.5
- ast       1 1555.5
- copper    1 1555.5
- edema     1 1557.1
- protime   1 1557.5
- albumin   1 1559.3
- age       1 1563.9
- stage     1 1564.0
- bili      1 1578.4

Step:  AIC=1551.54
Surv(timeYears, newStatus) ~ age + sex + ascites + hepato + stage +
   albumin + copper + ast + trig + bili + chol + edema + protime

         Df    AIC
- sex       1 1550.0
- chol      1 1550.1
- ascites   1 1550.5
```

```
- hepato   1 1551.1
- trig     1 1551.4
<none>       1551.5
- ast      1 1553.5
- copper   1 1553.7
- edema    1 1555.1
- protime  1 1555.6
- albumin  1 1557.3
- stage    1 1562.1
- age      1 1562.2
- bili     1 1576.4

Step:  AIC=1549.97
Surv(timeYears, newStatus) ~ age + ascites + hepato + stage +
    albumin + copper + ast + trig + bili + chol + edema + protime

         Df    AIC
- chol     1 1548.5
- ascites  1 1549.1
- hepato   1 1549.5
- trig     1 1549.9
<none>       1550.0
- ast      1 1552.1
- copper   1 1553.2
- edema    1 1553.2
- protime  1 1554.0
- albumin  1 1555.5
- stage    1 1560.4
- age      1 1562.6
- bili     1 1574.4

Step:  AIC=1548.49
Surv(timeYears, newStatus) ~ age + ascites + hepato + stage +
    albumin + copper + ast + trig + bili + edema + protime

         Df    AIC
- ascites  1 1547.5
- hepato   1 1548.0
- trig     1 1548.2
<none>       1548.5
- edema    1 1551.4
- ast      1 1551.4
- copper   1 1551.5
- protime  1 1552.4
- albumin  1 1554.2
- stage    1 1558.8
- age      1 1560.7
- bili     1 1576.5

Step:  AIC=1547.51
```

```
Surv(timeYears, newStatus) ~ age + hepato + stage + albumin +
    copper + ast + trig + bili + edema + protime

         Df    AIC
- trig     1 1546.5
- hepato   1 1546.8
<none>       1547.5
- ast      1 1550.1
- edema    1 1551.1
- protime  1 1551.9
- copper   1 1552.3
- albumin  1 1555.1
- stage    1 1558.5
- age      1 1561.2
- bili     1 1576.0

Step:  AIC=1546.55
Surv(timeYears, newStatus) ~ age + hepato + stage + albumin +
    copper + ast + bili + edema + protime

         Df    AIC
- hepato   1 1545.9
<none>       1546.5
- ast      1 1549.7
- edema    1 1550.7
- copper   1 1550.8
- protime  1 1551.1
- albumin  1 1553.7
- stage    1 1556.9
- age      1 1560.9
- bili     1 1575.5

Step:  AIC=1545.91
Surv(timeYears, newStatus) ~ age + stage + albumin + copper +
    ast + bili + edema + protime

         Df    AIC
<none>       1545.9
- ast      1 1549.0
- edema    1 1549.7
- copper   1 1550.5
- protime  1 1551.0
- albumin  1 1554.1
- stage    1 1559.4
- age      1 1560.1
- bili     1 1576.0
```