Colorectal cancer data: determinants of disease –free survival rates

Homework project - survival analysis with R

Data ScienceTech Institute,

4, rue de la Collégiale, 75005 Paris, France

Soütrik Banerjee

25th January, 2017

I. INTRODUCTION

Colorectal cancer (CRC) is a common and lethal disease. The risk of developing CRC is influenced by both environmental and genetic factors. CRC incidence and mortality rates vary markedly around the world. It accounts for over 9% of all cancer incidence. Globally, CRC is the third most diagnosed cancer in males and the second in females, with 1.4 million new cases and 6,94,000 deaths estimated to have occurred in 2012. The highest incidence rates are in Australia and New Zealand, Europe and North America; the lowest rates are found in China, India, and parts of Africa and South America.

These geographic differences appear to be attributable to differences in dietary and environmental exposures that are imposed upon a background of genetically determined susceptibility. Low socioeconomic status (SES) is also associated with an increased risk for the development of colorectal cancer. Potentially modifiable behaviors such as physical inactivity, unhealthy diet, smoking and obesity are thought to account for a substantial proportion (estimates of one-third to one-half) of the socioeconomic disparity in risk of new onset colorectal cancer. Other factors, particularly lower rates of CRC screening, also contribute substantively to SES differences in CRC risk.

In the current analysis, determinants of disease-free survival (DFS) was evaluated for cancer location in the gut, Duke's staging, gender, age at diagnosis, and if the patients received adjuvant radio- and chemo-therapy.

II. METHODS

The DFS was estimated in month's duration. Duke's staging had 3 stages in this data – "A", "B" and "C". The location of cancer diagnosis was – sigmoid colon, right colon, left colon and rectum. It was noted if the patients had received adjuvant radio- and/or chemo-therapy. Age at diagnosis was also noted. The event of interest was defined as the relapse or reappearance of CRC disease during the FUP. The time to event was also noted, and if the patient did not have any event during the FUP or death due to other causes than relapse or reappearance of the CRC disease (no competing risks analysis), the patient was considered as right censored.

Descriptive analysis for categorical variables was carried out by X^2 test, and by Fisher exact test when the expected cell count was less than 5. For continuous variables, t-test was carried out.

Kaplan-Meier (KM) curves (product-limit estimates) were estimated for the covariates; 95% confidence limits were estimated by the default Greenwood's method. Log-rank tests were employed to compare univariate associations for the different determinants (here, age was categorised by median split, with median included in the lower category).

Cox Proportional Hazard (PH) model was fitted with all covariates, using the default 'Efron' method for ties handling, and model assumptions tested.

Parametric survival models were fitted with the distributions: exponential, Weibull, log-normal, logistic, log-logistic and compared.

R software 3.3.1 (CRAN) was used for the statistical analysis.

III. RESULTS

The data consists of 226 patients, of which 120 males and 106 females, who had data for all the variables. Median follow-up period (FUP) for the patients enrolled was 38 months. Seventy-eight percent (n = 176) had an event during this FUP. Thirty-eight percent (n = 87) had received adjuvant chemo-therapy. Ten percent (n = 22) had received adjuvant radio-therapy, of which only one patient did not receive adjuvant chemo-therapy. Forty-five percent (n = 101) were located in the right colon and 41% (n = 93) were located in the left colon. Mean [SD] age at diagnosis was 66 [13] years. Eighteen percent (n = 41), 42% (n = 94) and 40% (n = 91) had Duke's staging "A", "B" and "C", respectively.

The median survival time was approximately 50 months (45 - 56 months, 95% CI).

Table 1. Summary statistics by gender.

	Female	Male	p	test
n =	106	120		
location (%)			0.057	exact
Rectum	11 (10.4)	19 (15.8)		
Colon	2 (1.9)	0 (0.0)		
Left	38 (35.8)	55 (45.8)		
Right	55 (51.9)	46 (38.3)		
dukes_stage (%)			0.970	
A	19 (17.9)	22 (18.3)		
B	45 (42.5)	49 (40.8)		
C	42 (39.6)	49 (40.8)		
age_diag (mean (sd))	67.84 (12.74)	64.43 (13.10)	0.049	
dfs_time (mean (sd))	42.43 (27.20)	44.48 (28.43)	0.583	
dfs_event (mean (sd))	0.79 (0.41)	0.77 (0.42)	0.643	
adjXRT = Y (%)	9 (8.5)	13 (10.8)	0.713	
adjCTX = Y (%)	40 (37.7)	47 (39.2)	0.933	

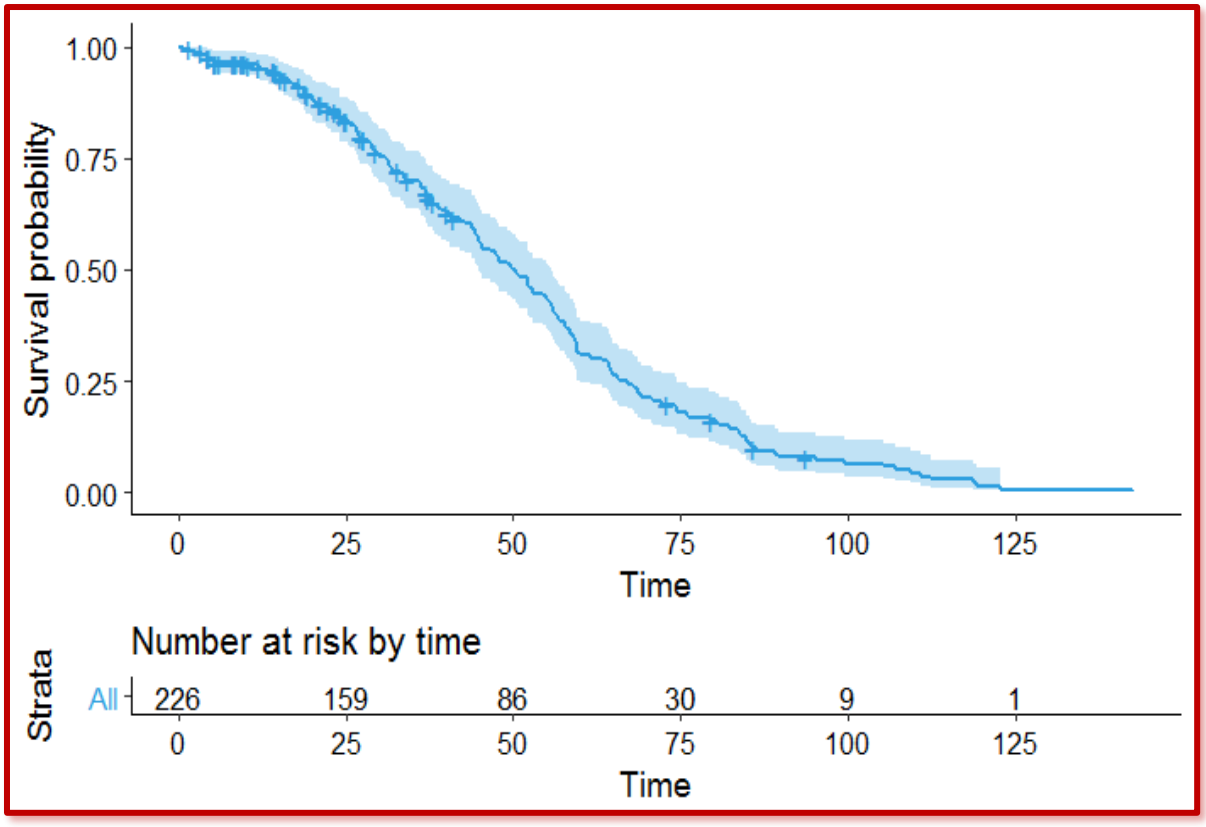

Figure 1. Kaplan-Meier plot for time to event (disease-free survival) in months.

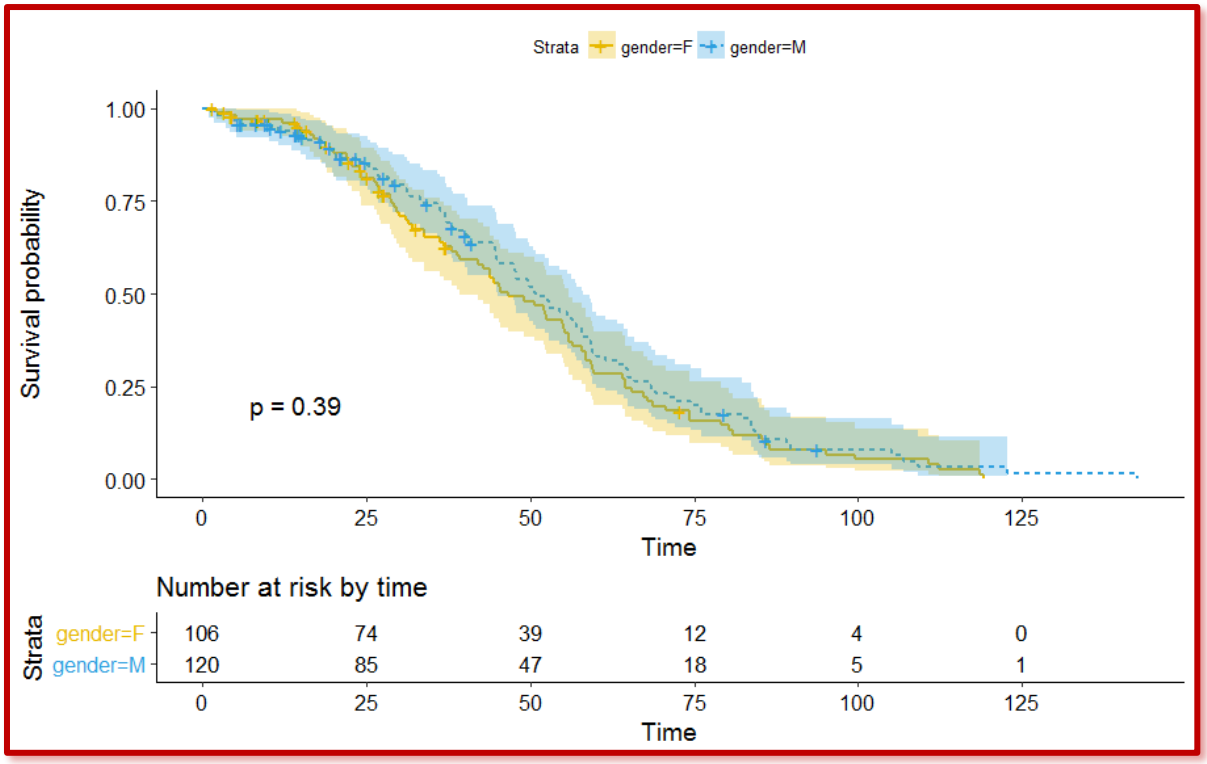

Figure 2. Kaplan-Meier plot for time to event (disease-free survival) in months, according to gender.

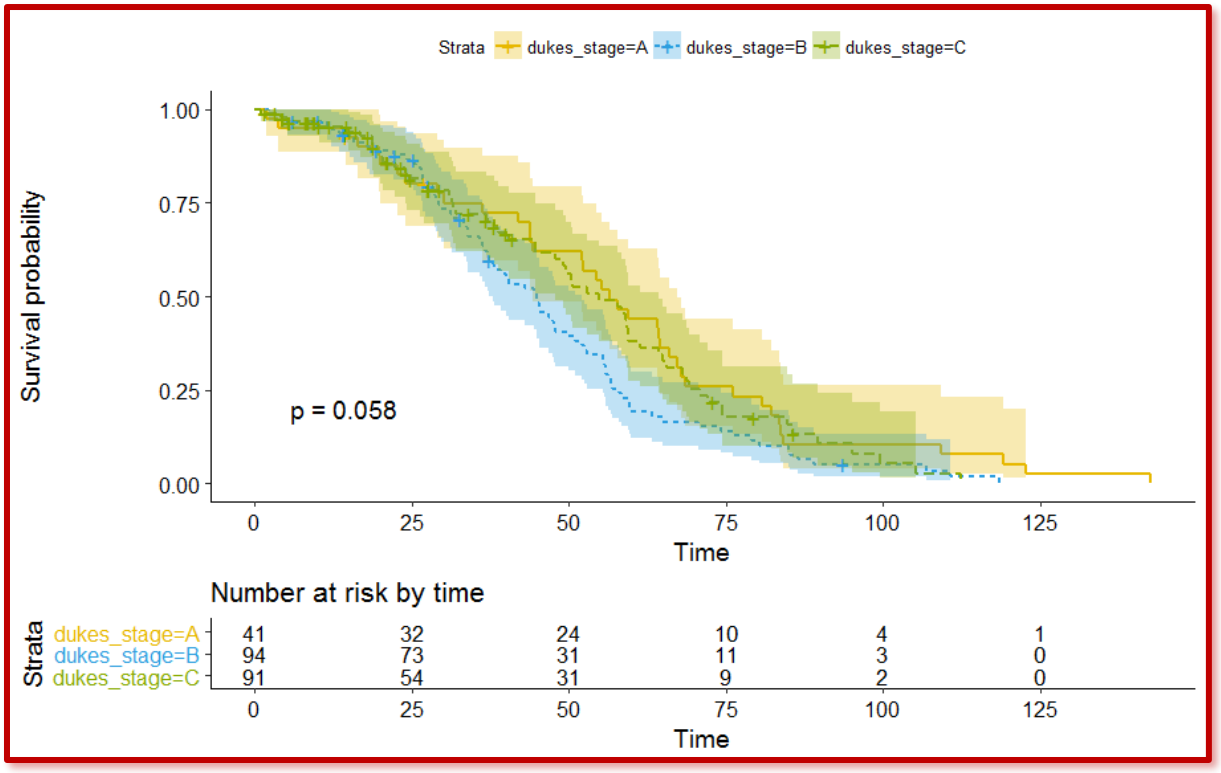

Figure 3. Kaplan-Meier plot for time to event (disease-free survival) in months, according to Duke's staging.

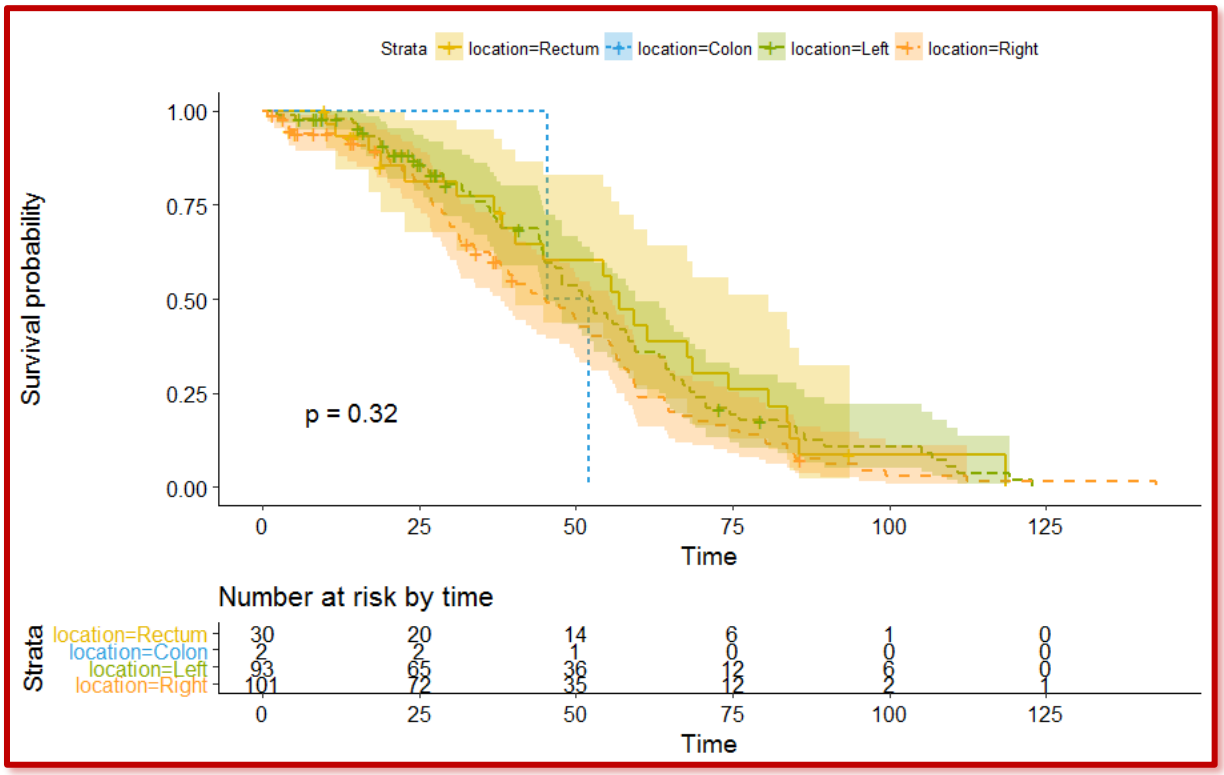

Figure 4. Kaplan-Meier plot for time to event (disease-free survival) in months, according to location.

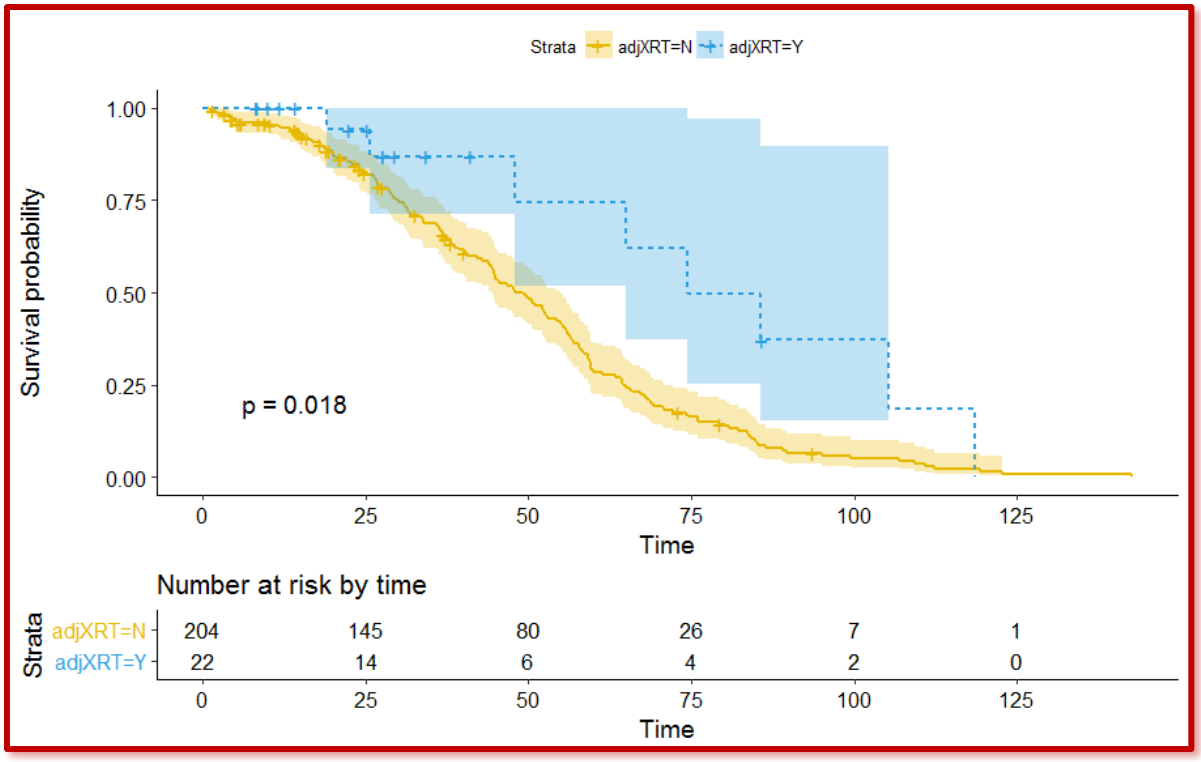

Figure 5. Kaplan-Meier plot for time to event (disease-free survival) in months, according to adjuvant radio-therapy (all adjuvant radio-therapy patients received adjuvant chemo-therapy also, except one patient).

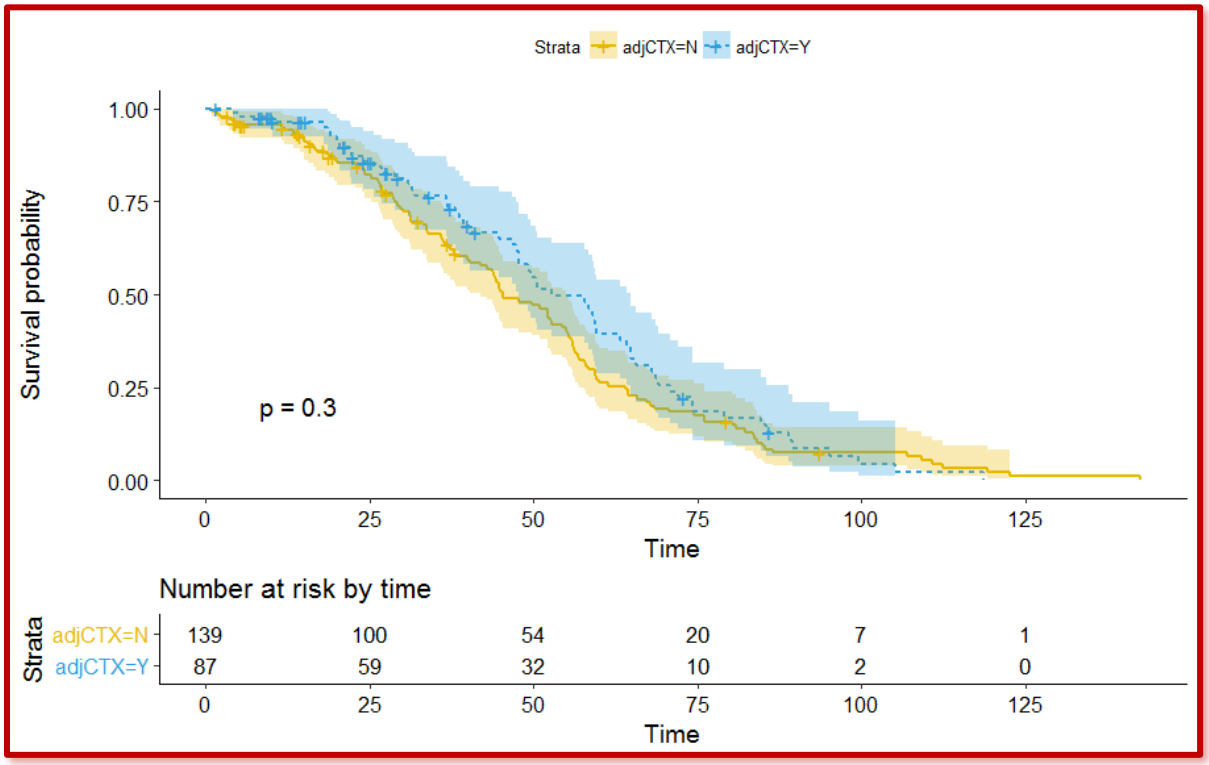

Figure 6. Kaplan-Meier plot for time to event (disease-free survival) in months, according to adjuvant chemo-therapy.

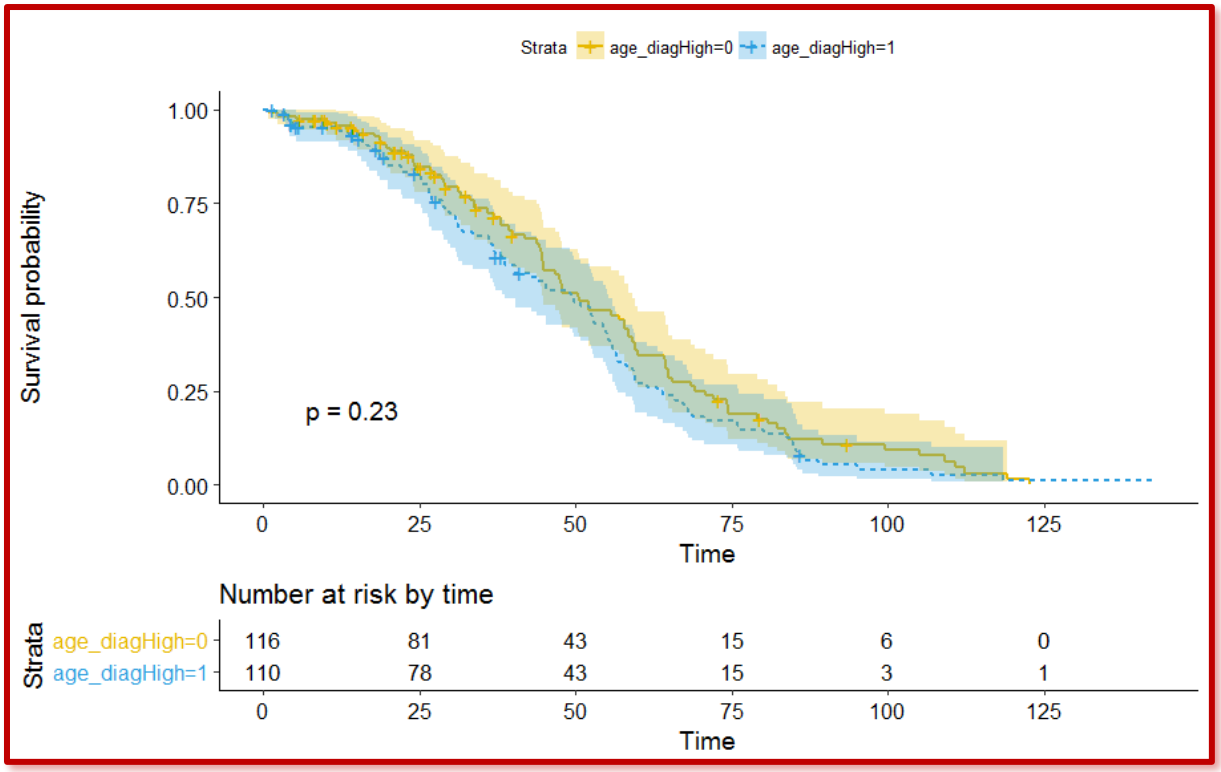

Figure 7. Kaplan-Meier plot for time to event (disease-free survival) in months, according to age at diagnosis (median split).

Cox PH model:

```
Call:
coxph(formula = Surv(dfs_time, dfs_event) ~ location + dukes_stage
+
    age_diagHigh + gender + adjXRT + adjCTX, data = clinical_data)

  n = 226, number of events = 176

                      coef     exp(coef)  se(exp(coef))   z
Pr(>|z|)
locationColon   0.37695    1.45782  0.75182     0.501    0.6161
locationLeft   -0.12257    0.88465  0.25214    -0.486    0.6269
locationRight   0.10420    1.10982  0.25532     0.408    0.6832
dukes_stageB    0.51019    1.66560  0.21377     2.387    0.0170 *
dukes_stageC    0.34113    1.40654  0.25875     1.318    0.1874
age_diagHigh1   0.08078    1.08413  0.16127     0.501    0.6164
```

genderM -0.06547 0.93663 0.15633 -0.419 0.6754

adjXRTY -0.87165 0.41826 0.40457 -2.155 **0.0312** *

adjCTXY -0.08826 0.91552 0.21169 -0.417 0.6767

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

	exp(coef)	exp(-coef)	lower .95	upper .95
locationColon	1.4578	0.6860	0.3340	6.3629
locationLeft	0.8846	1.1304	0.5397	1.4501
locationRight	1.1098	0.9010	0.6729	1.8305
dukes_stageB	1.6656	0.6004	**1.0955**	**2.5324**
dukes_stageC	1.4065	0.7110	0.8470	2.3356
age_diagHigh1	1.0841	0.9224	0.7903	1.4872
genderM	0.9366	1.0677	0.6894	1.2724
adjXRTY	0.4183	2.3908	**0.1893**	**0.9243**
adjCTXY	0.9155	1.0923	0.6046	1.3863

Concordance = 0.593 (se = 0.026)

Rsquare = 0.071 (max possible= 0.999)

Likelihood ratio test = 16.72 on 9 df, p = 0.05328

Wald test = 15.04 on 9 df, p = 0.08981

Score (logrank) test = 15.56 on 9 df, p = 0.07656

Cox PH model assumptions:

	rho	chisq	p
locationColon	0.060282	6.41e-01	0.423
locationLeft	-0.030695	1.75e-01	0.676
locationRight	-0.090844	1.57e+00	0.211
dukes_stageB	0.055269	5.89e-01	0.443
dukes_stageC	-0.008781	1.36e-02	0.907

```
age_diagHigh1 -0.000264 1.29e-05 0.997
genderM       -0.032214 1.86e-01 0.666
adjXRTY       -0.008110 1.18e-02 0.913
adjCTXY        0.126564 2.67e+00 0.102
GLOBAL               NA 7.05e+00 0.632
```

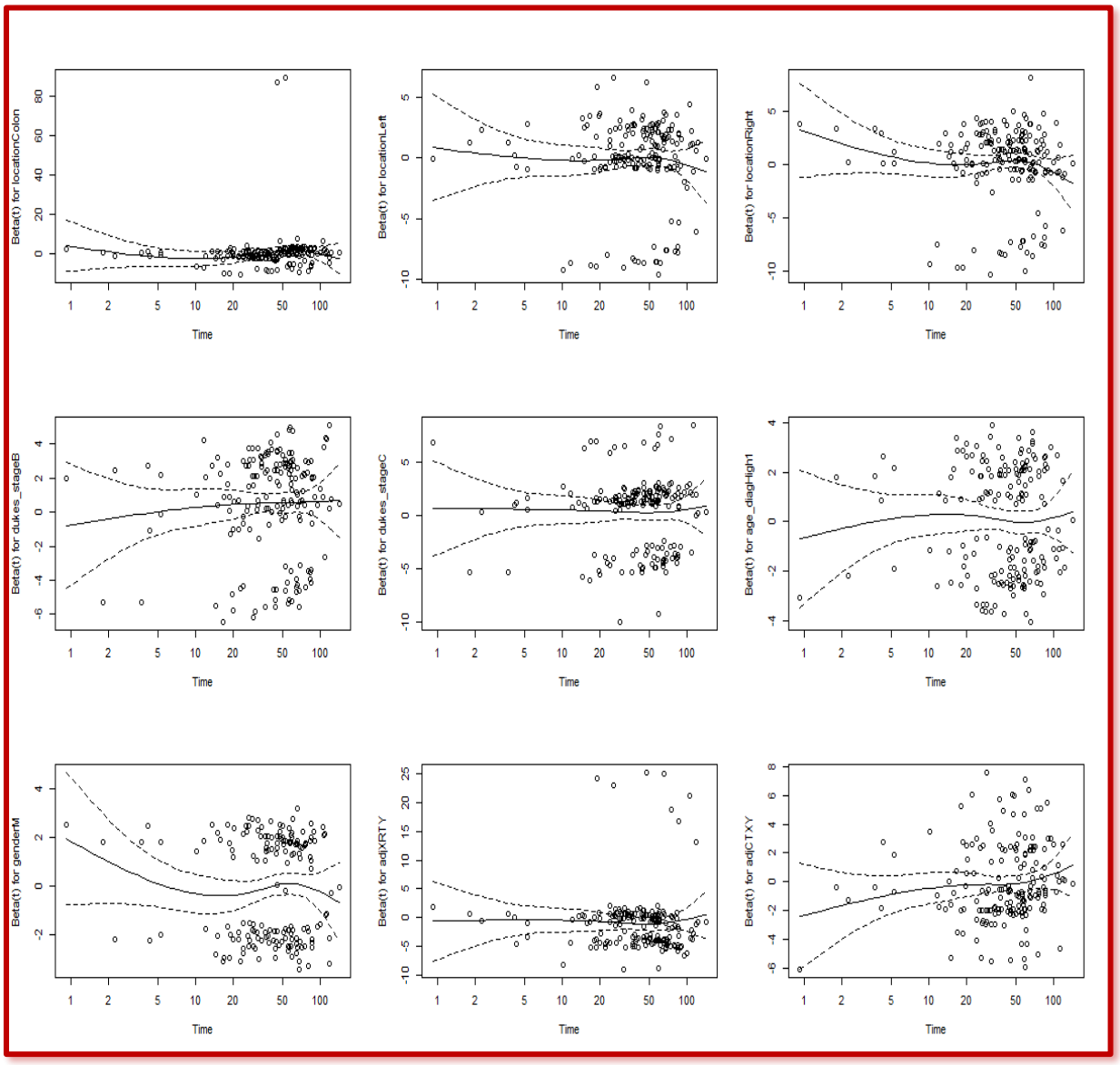

Figure 8. Schoenfeld's residuals plot with LOESS fit curve (95% CI).

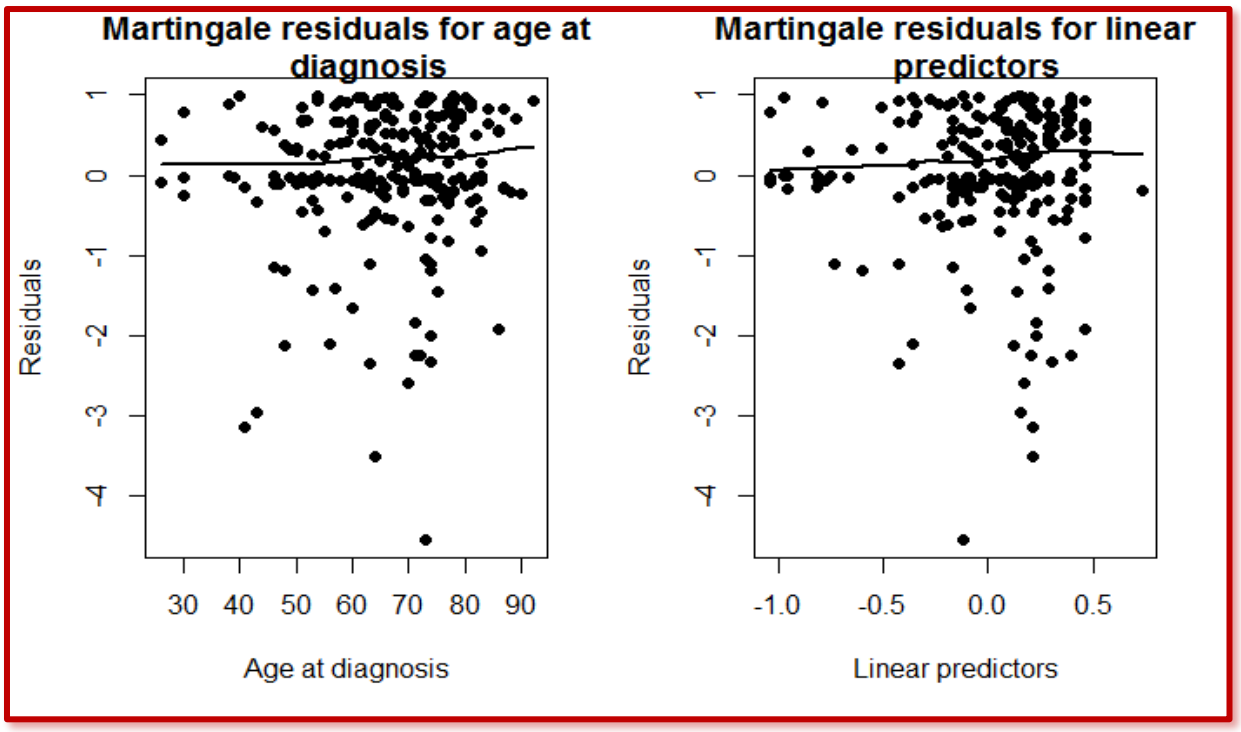

Figure 9. Martingale residuals for age at diagnosis and linear predictor.

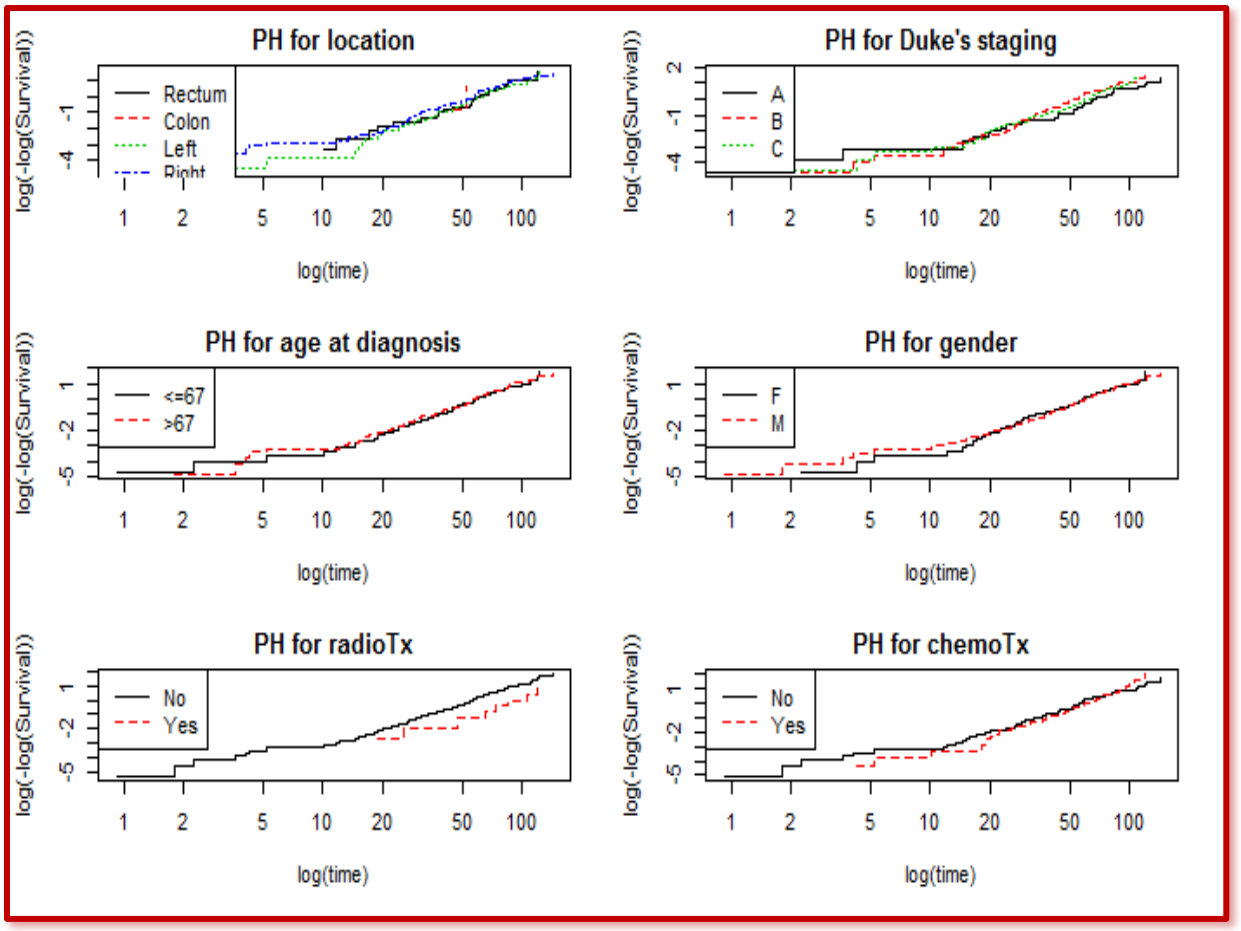

Figure 10. Plots for proportionality visual checks.

Prediction for Duke's staging and adjuvant radio-therapy:

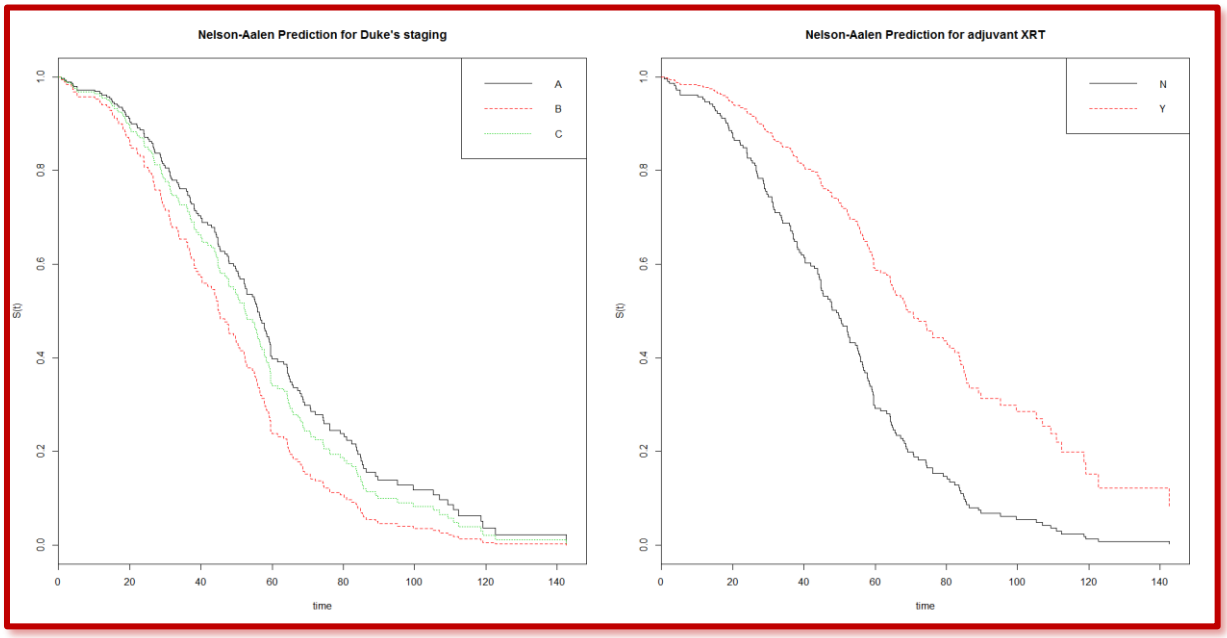

Figure 11. Nelson-Aalen predicted survival function for Duke's staging and adjuvant radio-therapy.

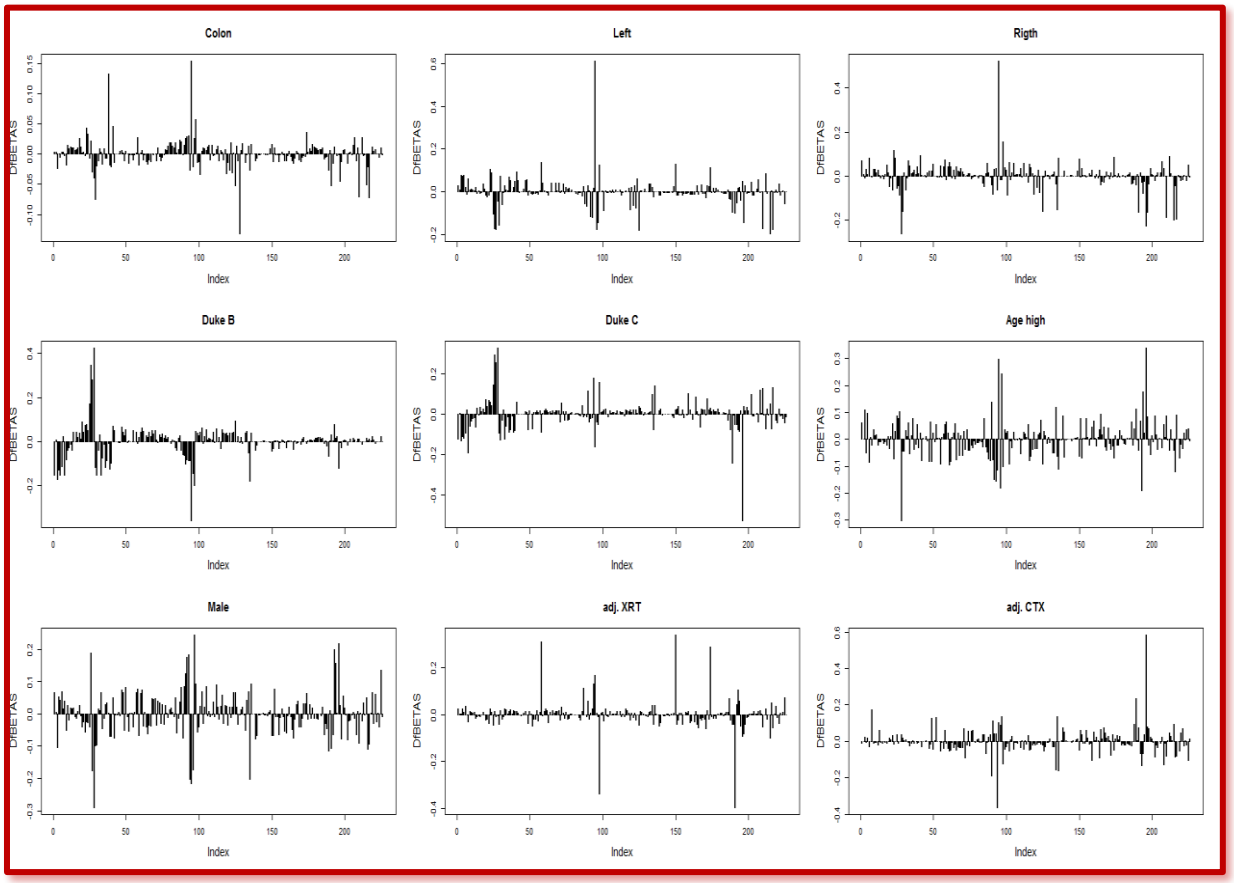

Figure 12. Influential observations on PH regression coefficients.

Parametric survival (Accelerated Failure Time) models:

Parametric models with the distributions: exponential, Weibull, log-normal, logistic, log-logistic and compared. The Weibull model has the least AIC, whose model fit parameters (in log scale) is given below.

```
Call:
survreg(formula = Surv(dfs_time, dfs_event) ~ location +
dukes_stage + age_diagHigh + gender + adjXRT + adjCTX, data =
clinical_data, dist = "weibull")
                Value Std. Error       z         p
(Intercept)     4.1816     0.1416  29.532 1.12e-191
locationColon -0.1513     0.3811  -0.397  6.91e-01
locationLeft   0.0559     0.1273   0.439  6.61e-01
locationRight -0.0536     0.1291  -0.415  6.78e-01
dukes_stageB  -0.2352     0.1042  -2.258  2.39e-02
dukes_stageC  -0.1594     0.1287  -1.239  2.15e-01
age_diagHigh1 -0.0466     0.0816  -0.571  5.68e-01
genderM        0.0312     0.0793   0.394  6.94e-01
adjXRTY        0.4410     0.2035   2.167  3.02e-02
adjCTXY        0.0368     0.1073   0.343  7.31e-01
Log(scale)    -0.6743     0.0581 -11.605  3.88e-31
Scale = 0.509
Weibull distribution
Loglik(model)  = -830.5   Loglik(intercept only)  = -838.4
Chisq= 15.73 on 9 degrees of freedom, p = 0.073
Number of Newton-Raphson Iterations: 7
n = 226
```

Parsimonious Weibull model:

The previously best fitted Weibull model was improved by fitting stepwise backward selection based on AIC, whose model parameters（in log scale） is given below.

Call:

survreg(formula = Surv(dfs_time, dfs_event) ~ dukes_stage + adjXRT, data = clinical_data, dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	4.178	0.0830	50.35	0.00e+00
dukes_stageB	−0.240	0.1004	−2.39	**1.69e−02**
dukes_stageC	−0.136	0.1078	−1.26	2.07e−01
adjXRTY	0.465	0.1880	2.47	**1.34e−02**
Log(scale)	−0.669	0.0581	−11.51	1.23e−30

Scale = **0.512**

Weibull distribution

Loglik(model) = −832 Loglik(intercept only) = −838.4

Chisq= 12.79 on 3 degrees of freedom, p= 0.0051

Number of Newton−Raphson Iterations: 7

n= 226

IV. DISCUSSION

In the present study, three different model fitting approaches were done to analyse the CRC data, namely, non-parametric KM curve, semi-parametric Cox PH model approach and finally, parametric model with exponential, Weibull, log-normal, logistic, log-logistic distributions. Of the predictors that were fitted to the model, having adjuvant radio-therapy (along with chemo-therapy) showed to improve DFS in CRC patients, and again Duke's staging B type showed worse outcome than staging A.

The Cox PH model was fitted with all predictor variables. The statistical significance increased for the Duke's staging of CRC. The individual covariates and globally the assumption of proportionality was not violated. This points towards the fact that perhaps no time varying (extended Cox model) nature of the covariates was significantly present, despite that the assumption of Cox proportionality is indeed very conservative. In addition, Schoenfeld (covariates) and Martingale residuals did not show any major pattern; for a time dependency to be present, Martingale residuals would have picked up a trend for the linear predictor (log-scale). No major influential covariables on the coefficients were detected by DfBetas.

Parametric survival modelling was carried out using different distributions of the survival times. Weibull model was found to be the best fit using AIC. A backward stepwise model using AIC was further employed to obtain a parsimonious model from the latter model. The scale factor was found to be equal to 0.512, which points towards an overall decrease in risk with time. In this model, Duke's stage B had HR ~ 1.27 compared to the full Cox model HR ~ 1.67, which means that Duke's B stage is associated with significantly worse outcome than Duke's A stage at all times. In the same way, getting adjuvant radio-therapy (with chemo-therapy) translated to a HR ~ 0.64 compared to the full Cox model HR ~ 0.42, which means taking anti-cancer treatment significantly delayed the reappearance of CRC.

A major limitation of the analysis was that competing risks were not considered. However, given that the disease reappearance would normally appear before death due to CRC as a major competing risk would not have significantly changed the outcome. However, premature death due to adverse reactions due to adjuvant therapy could have affected the outcome. This latter phenomenon could bias the result in favour of the adjuvant therapy. So in a future study, premature deaths due to competing risks need to evaluated in this cohort.

V. REFERENCES

1. Macrae FA. Colorectal cancer: Epidemiology, risk factors, and protective factors. http://www.uptodate.com/contents/colorectal-cancer-epidemiology-risk-factors-and-protective-factors.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;**65**(2):87.
3. Jemal A, Bray F, Center MM, et al. Global cancer statistics. *CA Cancer J Clin* 2011;**61**:69.
4. Doubeni CA, Laiyemo AO, Major JM, et al. Socioeconomic status and the risk of colorectal cancer: an analysis of more than a half million adults in the National Institutes of Health-AARP Diet and Health Study. *Cancer* 2012;**118**:3636.
5. Willett WC. Diet and cancer: an evolving picture. *JAMA* 2005;**293**:233.
6. Doubeni CA, Major JM, Laiyemo AO, et al. Contribution of behavioural risk factors and obesity to socioeconomic differences in colorectal cancer incidence. *J Natl Cancer Inst* 2012;**104**:1353.
7. Klabunde CN, Cronin KA, Breen N, et al. Trends in colorectal cancer test use among vulnerable populations in the United States. *Cancer Epidemiol Biomarkers Prev* 2011;**20**:1611.
8. R 3.3.1. Comprehensive R Archive Network. https://cran.r-project.org/.

I. APPENDIX OF R-CODES

```
---
title: "Survival analysis with colorectal cancer data"
author: "Soutrik Banerjee"
date: "24 January, 2017"
output:
  pdf_document: default
  word_document: default
---

## Import libraries
```{r load packages}
```
```

```
library(reshape2) # rename

library(tableone)

library(ggplot2)

library(dplyr)

library(sjmisc) # dicho()


library(survival)

library(survminer)

library(KMsurv)

library(survivalROC)
```

Import data

```{r Load data}

setwd("D:\\HP Pavilion D drive\\Mit\\DSTI\\Survival analysië met R")

load("ColorectalCaData.RData")

rm("gene_expression")
```

Summarise data

```{r Table 1}

summary1 <- CreateTableOne(data = clinical_data[, c(2:9)], strata = "gender")

print(summary1, exact = "location")
```

KM plots

```{r KM plots}
```

```
fit0 <- survfit(Surv(dfs_time, dfs_event) ~ 1, data =
clinical_data)

ggsurvplot(fit0, color = "#2E9FDF", conf.int = TRUE, break.time.by =
25,

           risk.table = TRUE)


fit.gender <- survfit(Surv(dfs_time, dfs_event) ~ gender, data =
clinical_data)

ggsurvplot(fit.gender, linetype = "strata", palette = c("#E7B800",
"#2E9FDF"),

           conf.int = TRUE, pval = TRUE, break.time.by = 25,
risk.table = TRUE,

           risk.table.y.text.col = TRUE)


fit.duke <- survfit(Surv(dfs_time, dfs_event) ~ dukes_stage, data =
clinical_data)

ggsurvplot(fit.duke, linetype = "strata", palette = c("#E7B800",
"#2E9FDF",

            "#86AA00"), conf.int = TRUE, pval = TRUE, break.time.by =
25,

           risk.table = TRUE, risk.table.y.text.col = TRUE)


fit.loc <- survfit(Surv(dfs_time, dfs_event) ~ location, data =
clinical_data)

ggsurvplot(fit.loc, linetype = "strata", palette = c("#E7B800",
"#2E9FDF",

            "#86AA00", "#FF9E29"), conf.int = TRUE, pval = TRUE,
break.time.by =

           25, risk.table = TRUE, risk.table.y.text.col = TRUE)
```

```
fit.xrt <- survfit(Surv(dfs_time, dfs_event) ~ adjXRT, data =
clinical_data)

ggsurvplot(fit.xrt, linetype = "strata", palette = c("#E7B800",
"#2E9FDF"),

           conf.int = TRUE, pval = TRUE, break.time.by = 25,
risk.table = TRUE,

           risk.table.y.text.col = TRUE)


fit.ctx <- survfit(Surv(dfs_time, dfs_event) ~ adjCTX, data =
clinical_data)

ggsurvplot(fit.ctx, linetype = "strata", palette = c("#E7B800",
"#2E9FDF"),

           conf.int = TRUE, pval = TRUE, break.time.by = 25,
risk.table = TRUE,

           risk.table.y.text.col = TRUE)


# dichotomise age by median split

clinical_data$age_diagHigh <- dicho(clinical_data$age_diag, dich.by
= "median",

                                    as.num = FALSE, var.label =
NULL, val.labels

                                    = NULL)

fit.age <- survfit(Surv(dfs_time, dfs_event) ~ age_diagHigh, data =
clinical_data)

ggsurvplot(fit.age, linetype = "strata", palette = c("#E7B800",
"#2E9FDF"),

           conf.int = TRUE, pval = TRUE, break.time.by = 25,
risk.table = TRUE,

           risk.table.y.text.col = TRUE)
```

Median (q50) survival time

```{r Median survival time}
survfit(Surv(dfs_time, dfs_event) ~ 1, data = clinical_data)
```

Cox PH regression

```{r Cox PH model}
mod1 <- coxph(Surv(dfs_time, dfs_event) ~ location + dukes_stage + age_diagHigh +
              gender + adjXRT + adjCTX, data = clinical_data) # Efron default
summary(mod1)
```

Cox PH model assumption

```{r Cox PH model check & Schoenfeld residuals}
mod1.zph <- cox.zph(mod1, transform = 'log')
mod1.zph

par(mfrow = c(3, 3))
plot(mod1.zph)
```

Cox PH plots in the full model

```{r strata(predictor.variable)}
loc.mod <- coxph(Surv(dfs_time, dfs_event) ~ strata(location) + dukes_stage +
                 age_diagHigh + gender + adjXRT + adjCTX, data = clinical_data)
```

```
duke.mod <- coxph(Surv(dfs_time, dfs_event) ~ location +
strata(dukes_stage) +
                    age_diagHigh + gender + adjXRT + adjCTX, data =
clinical_data)
age.mod <- coxph(Surv(dfs_time, dfs_event) ~ location + dukes_stage
+
                 strata(age_diagHigh) + gender + adjXRT + adjCTX,
                 data = clinical_data)
gen.mod <- coxph(Surv(dfs_time, dfs_event) ~ location + dukes_stage
+
                 age_diagHigh + strata(gender) + adjXRT + adjCTX,
                 data = clinical_data)
xrt.mod <- coxph(Surv(dfs_time, dfs_event) ~ location + dukes_stage
+
                 age_diagHigh + gender + strata(adjXRT) + adjCTX,
                 data = clinical_data)
ctx.mod <- coxph(Surv(dfs_time, dfs_event) ~ location + dukes_stage
+
                 age_diagHigh + gender + adjXRT + strata(adjCTX),
                 data = clinical_data)


par(mfrow = c(3, 2))
plot(survfit(loc.mod), fun = "cloglog", lty = 1:4, col = 1:4, ylab =
     "log(-log(Survival))", xlab = "log(time)", main = "PH for
location")
legend("topleft", legend = c("Rectum", "Colon", "Left", "Right"), lty
= 1:4, col = 1:4)
plot(survfit(duke.mod), fun = "cloglog", lty = 1:3, col = 1:3, ylab
=
```

```
     "log(-log(Survival))", xlab = "log(time)", main = "PH for
Duke's staging")

legend("topleft", legend = c("A", "B", "C"), lty = 1:3, col = 1:3)

plot(survfit(age.mod), fun = "cloglog", lty = 1:2, col = 1:2, ylab =

     "log(-log(Survival))", xlab = "log(time)", main = "PH for age
at diagnosis")

legend("topleft", legend = c("<=67", ">67"), lty = 1:2, col = 1:2)

plot(survfit(gen.mod), fun = "cloglog", lty = 1:2, col = 1:2, ylab =

     "log(-log(Survival))", xlab = "log(time)", main = "PH for
gender")

legend("topleft", legend = c("F", "M"), lty = 1:2, col = 1:2)

plot(survfit(xrt.mod), fun = "cloglog", lty = 1:2, col = 1:2, ylab =

     "log(-log(Survival))", xlab = "log(time)", main = "PH for
radioTx")

legend("topleft", legend = c("No", "Yes"), lty = 1:2, col = 1:2)

plot(survfit(ctx.mod), fun = "cloglog", lty = 1:2, col = 1:2, ylab =

     "log(-log(Survival))", xlab = "log(time)", main = "PH for
chemoTx")

legend("topleft", legend = c("No", "Yes"), lty = 1:2, col = 1:2)
```


Linearity of log(hazard) for continuous variables

```{r Martinagel residuals for age at diagnosis, linear.predictor}

resMart <- residuals(mod1, type = "martingale")


par(mfrow = c(1, 2))

plot(clinical_data$age_diag, resMart, main = "Martingale residuals
for age at
```

```
     diagnosis", xlab = "Age at diagnosis", ylab = "Residuals", pch =
16)

lines(lowess(clinical_data$age_diag, resMart), lwd = 2)

plot(mod1$linear.predictors, resMart, main = "Martingale residuals
for linear

     predictors", xlab = "Linear predictors", ylab = "Residuals", pch
= 16)

lines(lowess(mod1$linear.predictors, resMart), lwd = 2)
```

'Univariate' prediction Duke's staging, adjuvant XRT

```{r Duke's staging, adjuvant XRT}
mod2 <- coxph(Surv(dfs_time, dfs_event) ~ dukes_stage, data =
clinical_data)

mod2.pred <- survfit(mod2, newdata = data.frame(dukes_stage =

                     factor(c("A", "B", "C"), levels =

                     levels(clinical_data$dukes_stage)), type =
"aalen", se.fit =

                      TRUE))

mod3 <- coxph(Surv(dfs_time, dfs_event) ~ adjXRT, data =
clinical_data)

mod3.pred <- survfit(mod3, newdata = data.frame(adjXRT =

                     factor(c("N", "Y"), levels =

                     levels(clinical_data$adjXRT)), type = "aalen",
se.fit =

                     TRUE))


par(mfrow = c(1, 2))

plot(mod2.pred, lty = 1:3, col = 1:3, ylab = "S(t)", xlab = "time",
main =
```

```
     "Nelson-Aalen Prediction for Duke's staging")
legend("topright", legend = c("A", "B", "C"), lty = 1:3, col = 1:3)
plot(mod3.pred, lty = 1:2, col = 1:2, ylab = "S(t)", xlab = "time",
main =
     "Nelson-Aalen Prediction for adjuvant XRT")
legend("topright", legend = c("N", "Y"), lty = 1:2, col = 1:2)
```

Influential observations

```{r dfbetas}
dfbetas <- residuals(mod1, type = "dfbetas")

par(mfrow = c(3, 3), cex.main = 1.4, cex.lab = 1.4)
plot(dfbetas[, 1], type = "h", main = "Colon", ylab = "DfBETAS", lwd
= 2)
plot(dfbetas[, 2], type = "h", main = "Left", ylab = "DfBETAS", lwd =
2)
plot(dfbetas[, 3], type = "h", main = "Rigth", ylab = "DfBETAS", lwd
= 2)
plot(dfbetas[, 4], type = "h", main = "Duke B", ylab = "DfBETAS", lwd
= 2)
plot(dfbetas[, 5], type = "h", main = "Duke C", ylab = "DfBETAS", lwd
= 2)
plot(dfbetas[, 6], type = "h", main = "Age high", ylab = "DfBETAS",
lwd = 2)
plot(dfbetas[, 7], type = "h", main = "Male", ylab = "DfBETAS", lwd =
2)
plot(dfbetas[, 8], type = "h", main = "adj. XRT", ylab = "DfBETAS",
lwd = 2)
```

```
plot(dfbetas[, 9], type = "h", main = "adj. CTX", ylab = "DfBETAS",
lwd = 2)
```

Parametric survival analysis

```{r Parametric models}
mod.expo <- survreg(Surv(dfs_time, dfs_event) ~ location +
dukes_stage +
                    age_diagHigh + gender + adjXRT + adjCTX,
dist="exponential",
                    data = clinical_data)
mod.weib <- survreg(Surv(dfs_time, dfs_event) ~ location +
dukes_stage +
                    age_diagHigh + gender + adjXRT + adjCTX,
dist="weibull",
                    data = clinical_data)
mod.lnorm <- survreg(Surv(dfs_time, dfs_event) ~ location +
dukes_stage +
                    age_diagHigh + gender + adjXRT + adjCTX,
dist="lognormal",
                    data = clinical_data)
mod.logit <- survreg(Surv(dfs_time, dfs_event) ~ location +
dukes_stage +
                    age_diagHigh + gender + adjXRT + adjCTX,
dist="logistic",
                    data = clinical_data)
mod.llogit <- survreg(Surv(dfs_time, dfs_event) ~ location +
dukes_stage +
                    age_diagHigh + gender + adjXRT + adjCTX,
dist="loglogistic",
```

```
                    data = clinical_data)


# compare models
param.mod <- list(mod.expo, mod.weib, mod.lnorm, mod.logit,
mod.llogit)
lapply(param.mod, summary)
sapply(param.mod, AIC) # the lower, the better


# parsimonious model
mod.aic <- stepAIC(mod.weib, direction = "backward") # backward
stepwise default
summary(mod.aic)
```


Survival ROC for continuous variable
```{r Survival ROCfor age at diagnosis}
age.ROC <- survivalROC(Stime = clinical_data$dfs_time, status =
                         clinical_data$dfs_event, marker =
clinical_data$age_diag,
                       predict.time = 125, method = "KM")
age.ROC # AUC ~ 0.4 !!!
```