Survival Analysis of Mortality of Adjuvant Chemotherapy for Colon Cancer

Yijia Jiang, Ziyan Xu

2022-11-27

1. Data import and Examination

Recall that there are two records for each patient indicated by the event type (etype) variable, where etype == 1 refers to the event of a recurrence and etype == 2 indicates death. In order to answer our first research question, which is to study the time until death, we must create a marginal model by subsetting the colon data to only include the event of mortality. To get an overview of the mortality subset we use the survfit function and plot the Kaplan-Meier Estimate between the three different treatments.

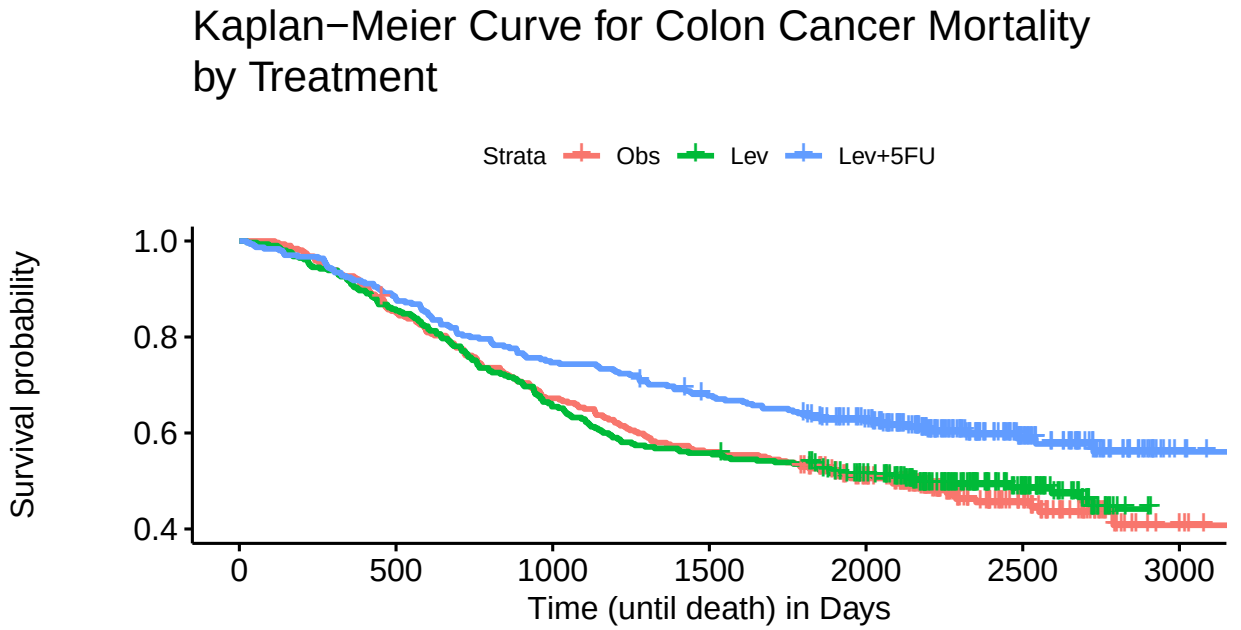
```
# import the dataset from survival package
data(cancer, package = "survival")
colon <- as_tibble(colon)

# subset death data
colon.death <- subset(colon, etype == 2)
```

2. Kaplan-Meier Survival Estimate

```
death.fit <- survfit(Surv(time,status) ~ rx, data = colon.death)

ggsurvplot(death.fit, conf.int = F, break.time.by = 500, ylim = c(0.4,1.0),
           font.x.size = 12, font.y.size = 12, font.legend.size = 9, legend.labs = c("Obs","Lev","Lev+5
           title = "Kaplan-Meier Curve for Colon Cancer Mortality \nby Treatment",
           xlab = "Time (until death) in Days",
           risk.table = T, risk.table.height = 0.25, risk.table.fontsize = 4,
           tables.theme = theme_cleantable())
```



Number at risk

Obs	315	267	211	176	141	50	6
Lev	310	265	203	173	145	58	4
Lev+5FU	304	267	227	203	170	65	7

From the plot above, there is some indication that patients who received the adjuvant treatment with levamisole plus fluorouracil (Lev+5Fu) have a higher survival probability than patients with no further treatment and patients who received the treatment with levamisole alone.

3. Log-Rank Test

We do a proper Log-rank hypothesis test to test the null hypothesis of no difference among the three treatments in the mortality model.

```
d.rx.coxph <- coxph(Surv(time, status) ~rx, data = colon.death)
summary(d.rx.coxph)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon.death)
##
##   n= 929, number of events= 452
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.02664   0.97371  0.11030 -0.241  0.80917
## rxLev+5FU -0.37171   0.68955  0.11875 -3.130  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
```

```
## rxLev        0.9737      1.027    0.7844    1.2087
## rxLev+5FU    0.6896      1.450    0.5464    0.8703
##
## Concordance= 0.536  (se = 0.013 )
## Likelihood ratio test= 12.15  on 2 df,   p=0.002
## Wald test            = 11.56  on 2 df,   p=0.003
## Score (logrank) test = 11.68  on 2 df,   p=0.003
```

From this log-rank test, we get a p-value that is closed to 0.002, which is significant at a 0.05 level. We want to conclude that there is a significant difference among the three treatments in the mortality model.

4. Tidy the Data for Model Development

Moreover, we notice that variables nodes and node4 both indicate similar information regarding the amount of positive lymph nodes an individual has. The variable nodes measures the number of lymph nodes with detectable cancer while node4 indicates whether there are more than 4 positive lymph nodes (0 = No, 1 =Yes). Therefore, we decided to use only the variable node4 in our analysis.

Additionally, it is important to note that there are columns that contain NA values. Out of 929 observations, 41 of them contain NA values in at least one column. Since observations that contain NA values make up only 4.41% of our data, we decided that removing them wouldn't cause a big effect on the variable selection process. By removing observations with NA values, created a new mortality dataset colon.death1 which contains 888 observations.

```
# check for missing values
apply(is.na(colon.death), 2, which) %>%
  str()
```

```
## List of 16
##  $ id      : int(0)
##  $ study   : int(0)
##  $ rx      : int(0)
##  $ sex     : int(0)
##  $ age     : int(0)
##  $ obstruct: int(0)
##  $ perfor  : int(0)
##  $ adhere  : int(0)
##  $ nodes   : int [1:18] 94 99 143 189 199 338 358 365 383 502 ...
##  $ status  : int(0)
##  $ differ  : int [1:23] 64 83 90 161 190 200 202 244 294 325 ...
##  $ extent  : int(0)
##  $ surg    : int(0)
##  $ node4   : int(0)
##  $ time    : int(0)
##  $ etype   : int(0)
```

```
# remove missing values
colon.death1 <- na.omit(colon.death)
```

5. Cox PH Model

5.1 Model Selection

We now use forward selection with Bayesian information criterion (BIC) to determine the covariates that best represent an appropriate cox proportional hazards model for the event of death. Within each step, we chose the model that has the lowest AIC and BIC value.

```
# fit all the variables in the model
d.model.full <- coxph(Surv(time, status) ~ sex + age + obstruct + perfor + adhere + differ + extent + su

# stepwise selection with BIC criterion
d.model.bic <- step(d.model.full, direction = "both", k = log(nrow(colon.death1)))
```

```
## Start:  AIC=5474.96
## Surv(time, status) ~ sex + age + obstruct + perfor + adhere +
##     differ + extent + surg + node4
##
##            Df    AIC
## - perfor    1 5468.2
## - sex       1 5468.3
## - differ    1 5470.0
## - adhere    1 5470.0
## - age       1 5471.1
## - obstruct  1 5473.2
## - surg      1 5473.3
## <none>        5475.0
## - extent    1 5484.4
## - node4     1 5547.8
##
## Step:  AIC=5468.17
## Surv(time, status) ~ sex + age + obstruct + adhere + differ +
##     extent + surg + node4
##
##            Df    AIC
## - sex       1 5461.5
## - differ    1 5463.2
## - adhere    1 5463.3
## - age       1 5464.3
## - obstruct  1 5466.5
## - surg      1 5466.5
## <none>        5468.2
## + perfor    1 5475.0
## - extent    1 5477.7
## - node4     1 5541.2
##
## Step:  AIC=5461.49
## Surv(time, status) ~ age + obstruct + adhere + differ + extent +
##     surg + node4
##
##            Df    AIC
## - differ    1 5456.6
## - adhere    1 5456.6
```

```
## - age       1 5457.6
## - obstruct  1 5459.8
## - surg      1 5459.9
## <none>        5461.5
## + sex       1 5468.2
## + perfor    1 5468.3
## - extent    1 5471.0
## - node4     1 5534.4
##
## Step:  AIC=5456.57
## Surv(time, status) ~ age + obstruct + adhere + extent + surg +
##     node4
##
##            Df    AIC
## - adhere    1 5452.2
## - age       1 5452.8
## - obstruct  1 5454.6
## - surg      1 5455.0
## <none>        5456.6
## + differ    1 5461.5
## + sex       1 5463.2
## + perfor    1 5463.3
## - extent    1 5467.4
## - node4     1 5533.3
##
## Step:  AIC=5452.15
## Surv(time, status) ~ age + obstruct + extent + surg + node4
##
##            Df    AIC
## - age       1 5449.0
## - obstruct  1 5450.1
## - surg      1 5450.7
## <none>        5452.2
## + adhere    1 5456.6
## + differ    1 5456.6
## + sex       1 5458.8
## + perfor    1 5458.8
## - extent    1 5464.4
## - node4     1 5528.7
##
## Step:  AIC=5449
## Surv(time, status) ~ obstruct + extent + surg + node4
##
##            Df    AIC
## - obstruct  1 5446.1
## - surg      1 5447.8
## <none>        5449.0
## + age       1 5452.2
## + adhere    1 5452.8
## + differ    1 5453.2
## + sex       1 5455.7
## + perfor    1 5455.7
## - extent    1 5461.4
## - node4     1 5522.8
```

```
##
## Step:  AIC=5446.11
## Surv(time, status) ~ extent + surg + node4
##
##            Df    AIC
## - surg      1 5445.0
## <none>        5446.1
## + obstruct  1 5449.0
## + adhere    1 5450.1
## + age       1 5450.1
## + differ    1 5450.6
## + perfor    1 5452.7
## + sex       1 5452.9
## - extent    1 5460.3
## - node4     1 5518.6
##
## Step:  AIC=5445.04
## Surv(time, status) ~ extent + node4
##
##            Df    AIC
## <none>        5445.0
## + surg      1 5446.1
## + obstruct  1 5447.8
## + adhere    1 5448.8
## + age       1 5448.9
## + differ    1 5449.5
## + perfor    1 5451.6
## + sex       1 5451.8
## - extent    1 5458.9
## - node4     1 5515.7
```

```
# stepwise selection with AIC criterion
d.model.aic <- step(d.model.full, direction = "both", k = 2)
```

```
## Start:  AIC=5431.86
## Surv(time, status) ~ sex + age + obstruct + perfor + adhere +
##     differ + extent + surg + node4
##
##            Df    AIC
## - perfor    1 5429.9
## - sex       1 5430.0
## - differ    1 5431.7
## - adhere    1 5431.7
## <none>        5431.9
## - age       1 5432.8
## - obstruct  1 5434.9
## - surg      1 5435.0
## - extent    1 5446.1
## - node4     1 5509.5
##
## Step:  AIC=5429.86
## Surv(time, status) ~ sex + age + obstruct + adhere + differ +
##     extent + surg + node4
##
```

```
##            Df    AIC
## - sex       1 5428.0
## - differ    1 5429.7
## - adhere    1 5429.8
## <none>        5429.9
## - age       1 5430.8
## + perfor    1 5431.9
## - obstruct  1 5433.0
## - surg      1 5433.0
## - extent    1 5444.2
## - node4     1 5507.7
##
## Step:  AIC=5427.96
## Surv(time, status) ~ age + obstruct + adhere + differ + extent +
##     surg + node4
##
##            Df    AIC
## - differ    1 5427.8
## - adhere    1 5427.9
## <none>        5428.0
## - age       1 5428.9
## + sex       1 5429.9
## + perfor    1 5430.0
## - obstruct  1 5431.0
## - surg      1 5431.2
## - extent    1 5442.3
## - node4     1 5505.7
##
## Step:  AIC=5427.83
## Surv(time, status) ~ age + obstruct + adhere + extent + surg +
##     node4
##
##            Df    AIC
## <none>        5427.8
## + differ    1 5428.0
## - adhere    1 5428.2
## - age       1 5428.9
## + sex       1 5429.7
## + perfor    1 5429.8
## - obstruct  1 5430.7
## - surg      1 5431.1
## - extent    1 5443.5
## - node4     1 5509.3
```

The resulting model with the lowest BIC is: Surv(time, status) ~ extent + node4

The resulting model with the lowest AIC is: Surv(time, status) ~ age + obstruct + adhere + extent + surg + node4

Next, we used the Analysis of Deviance procedure to get the proper Likelihood Ratio Test to confirm if each of the covariates selected by the forward selection method is significant to include in the Cox Proportional Model.

```
anova(d.model.bic)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##         loglik  Chisq Df Pr(>|Chi|)
## NULL   -2767.9
## extent -2754.5 26.899  1  2.143e-07 ***
## node4  -2715.7 77.461  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

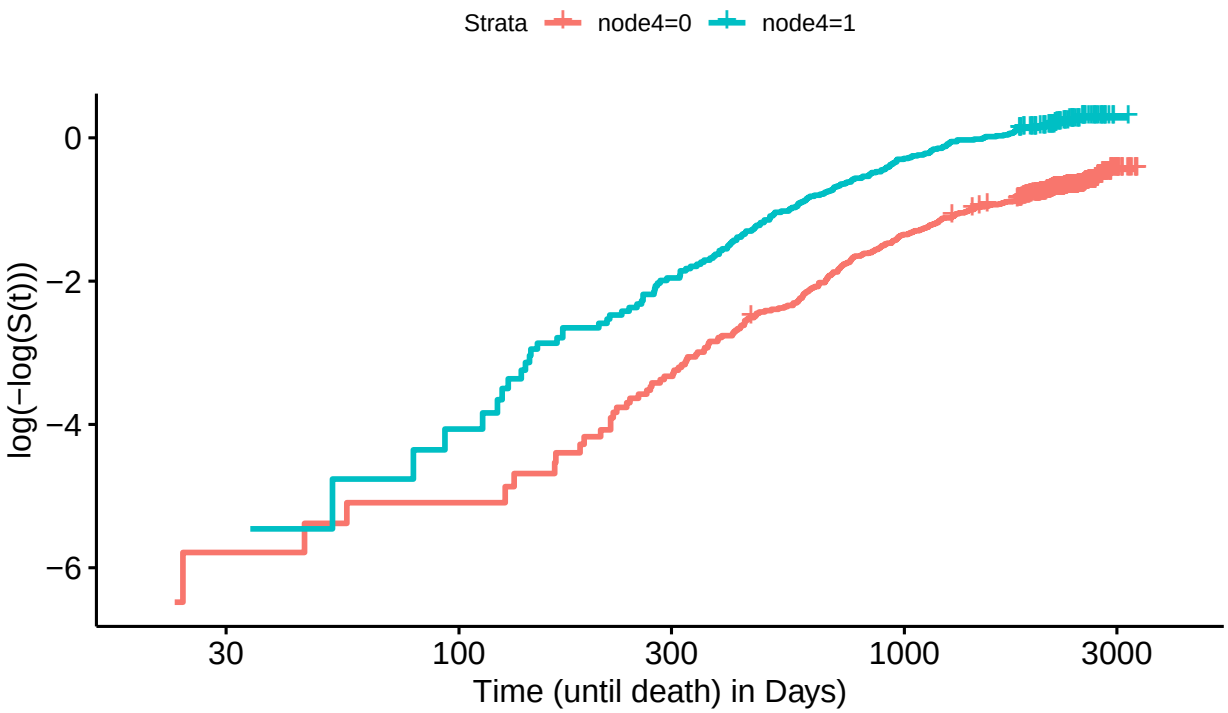
We can see that p-values for covariates extent and node4 are much smaller than 0.05, indicating that they have a significant effect on time until death. Therefore, we will include these four covariates in our Cox PH model.

5.2 Model Diagnostic

5.2.1 Log of Negative Log of Estimated Survival Function

To check the proportional hazards assumption for this model, we use a diagnostic plot such as the log of negative log of estimated survival function. Then we checked the significance of each covariate again using Analysis of Deviance procedure to ensure that our previous model is still valid. First comes to covariate node4.

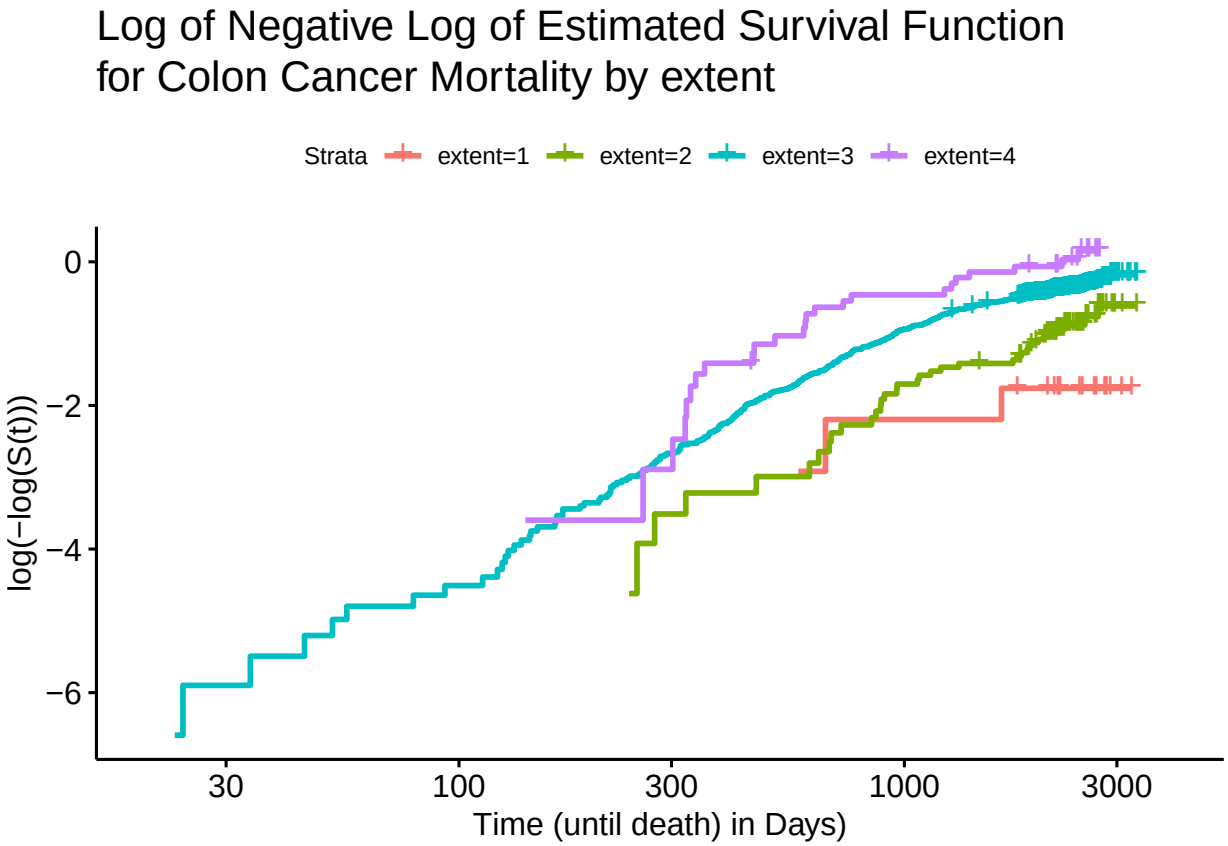
```
d.node4.fit <- survfit(Surv(time, status) ~ node4, data = colon.death1)
ggsurvplot(d.node4.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by
```



According to the plot, the two curves in this C-log-log plot cross over at the beginning of the study but appear to be parallel to each other after 100 days. Since the data is oftentimes noisy at the beginning of the study, the cross over does not cause too much concern. Overall, we believe that the cox proportional assumption is appropriate for the covariate node4 since the curves are consistently parallel throughout most of the study.

We continue to plot the C-log-log plot for the covariate extent.

```
d.extent.fit <- survfit(Surv(time, status) ~ extent, data = colon.death1)
ggsurvplot(d.extent.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by
```



According to the plot, the curves in the C-log-log plot are crossing over after 100 days. Since there are not enough data points in each extent group to show a more comprehensive trend, it's hard for us to make a decision based on the plot.

5.2.2 Test Interaction for Proportionality

```
d.model.bic.inter <- coxph(Surv(time, status) ~ extent + node4 + extent * log(time) + node4 * log(time)
summary(d.model.bic.inter)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ extent + node4 + extent *
##     log(time) + node4 * log(time), data = colon.death1)
##
##   n= 888, number of events= 430
##
##                       coef  exp(coef)   se(coef)       z Pr(>|z|)
## extent           -3.689e-01  6.915e-01  1.499e-01  -2.461   0.0138 *
## node4            -9.444e-01  3.889e-01  1.331e-01  -7.096 1.29e-12 ***
## log(time)        -1.723e+02  1.458e-75  9.366e+00 -18.398  < 2e-16 ***
## extent:log(time)  5.105e-02  1.052e+00  2.173e-02   2.349   0.0188 *
## node4:log(time)   1.480e-01  1.159e+00  1.976e-02   7.487 7.07e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## extent            6.915e-01  1.446e+00 5.154e-01 9.276e-01
## node4             3.889e-01  2.571e+00 2.996e-01 5.048e-01
## log(time)         1.458e-75  6.857e+74 1.553e-83 1.369e-67
## extent:log(time) 1.052e+00  9.502e-01 1.008e+00 1.098e+00
## node4:log(time)  1.159e+00  8.625e-01 1.115e+00 1.205e+00
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 4892  on 5 df,   p=<2e-16
## Wald test            = 456.4  on 5 df,   p=<2e-16
## Score (logrank) test = 2470  on 5 df,   p=<2e-16
```

Based on the test interaction for proportionality, the result shows that the interaction of `node4` with `log(time)` and the interaction of `extent` with `log(time)` are both significant (i.e., less than 0.05). Therefore, we can conclude that the proportionality assumptions for `node4` and `exntent` are met.

5.2.3 Goodness of fit test

```
cox.zph(coxph(formula = Surv(time, status) ~ node4 + extent, data = colon.death1))
```

```
##        chisq df      p
## node4   6.52  1 0.0107
## extent  4.18  1 0.0408
## GLOBAL 10.00  2 0.0067
```