Survival Analysis of Recurrence of Adjuvant Chemotherapy for Colon Cancer

Kaitlyn Wang

2022-11-30

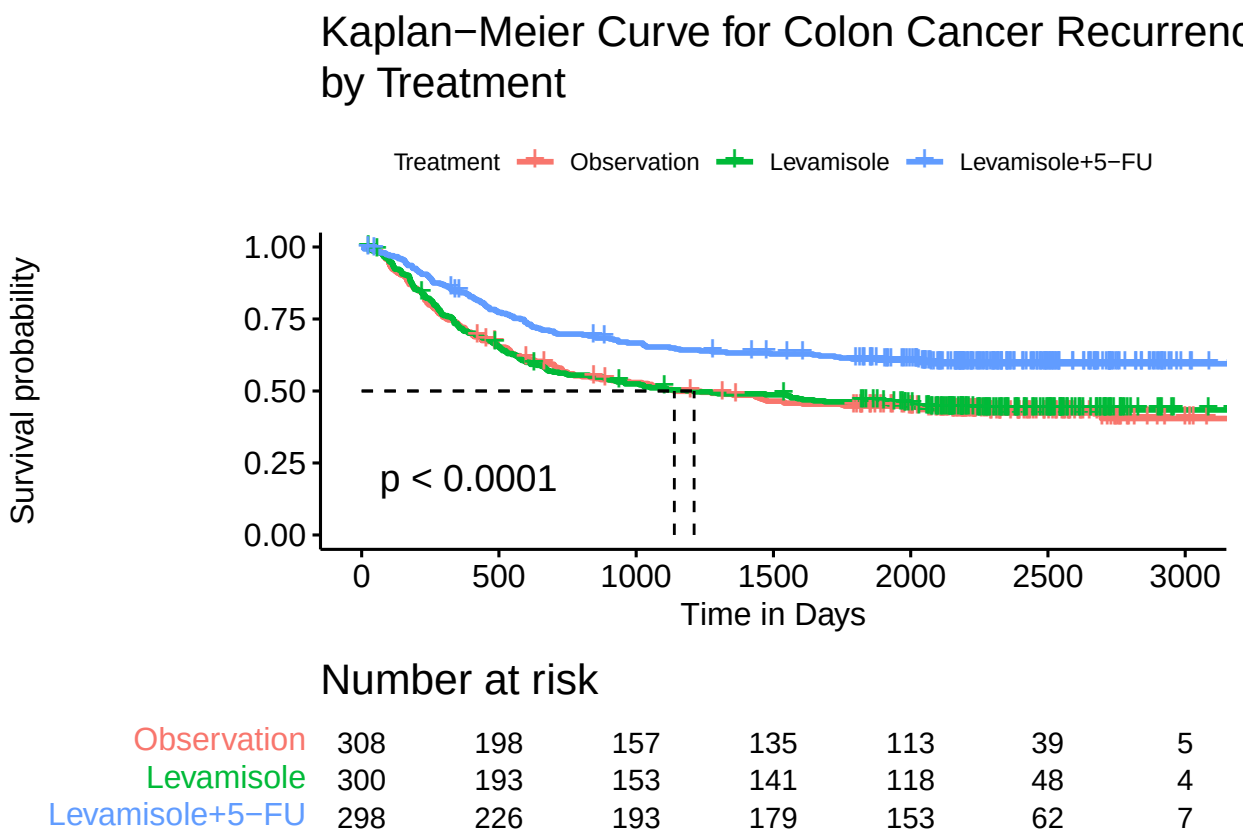
1. Data Import and Data Cleaning

```
colon <- survival::colon
colon.recur <- colon %>% filter(etype == '1') %>%
               dplyr::select(-c(nodes,id,study,etype)) %>%
               mutate(rx = as.factor(as.numeric(rx)),
                      sex = as.factor(sex),
                      obstruct = as.factor(obstruct),
                      perfor = as.factor(perfor),
                      adhere = as.factor(adhere),
                      differ = as.factor(differ),
                      extent = as.factor(extent),
                      surg = as.factor(surg),
                      node4 =as.factor(node4)) %>%
               drop_na()
```

2. Kaplan-Meier Survival Estimate

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

ggsurvplot(recur.fit, conf.int = F, break.time.by = 500, pval = TRUE,
           font.x.size = 12, font.y.size = 12, font.legend.size = 9, surv.median.line = "hv",
           legend.title = "Treatment", legend.labs = c("Observation", "Levamisole","Levamisole+5-FU"),
           title = "Kaplan-Meier Curve for Colon Cancer Recurrence \nby Treatment",
           xlab = "Time in Days",
           risk.table = T, risk.table.height = 0.25, risk.table.fontsize = 4,
           tables.theme = theme_cleantable())
```



3. Log-Rank Test

```
coxph(Surv(time, status) ~ rx, data = colon.recur)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon.recur)
##
##         coef exp(coef) se(coef)      z        p
## rx2 -0.02393   0.97636  0.10850 -0.221    0.825
## rx3 -0.50379   0.60424  0.11934 -4.222 2.43e-05
##
## Likelihood ratio test=22.81  on 2 df, p=1.113e-05
## n= 906, number of events= 458
```

```
summary(coxph(Surv(time, status) ~ rx, data = colon.recur))
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon.recur)
##
##   n= 906, number of events= 458
##
##         coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## rx2 -0.02393   0.97636  0.10850 -0.221    0.825
## rx3 -0.50379   0.60424  0.11934 -4.222 2.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## rx2    0.9764      1.024    0.7893    1.2077
## rx3    0.6042      1.655    0.4782    0.7635
##
## Concordance= 0.554  (se = 0.013 )
## Likelihood ratio test= 22.81  on 2 df,   p=1e-05
## Wald test            = 21.23  on 2 df,   p=2e-05
## Score (logrank) test = 21.66  on 2 df,   p=2e-05
```

4. Cox PH Model

4.1 Model Selection

```r
# full model
r.model.full <- coxph(Surv(time, status) ~ ., data = colon.recur)

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

```
## Start:  AIC=5815.3
## Surv(time, status) ~ rx + sex + age + obstruct + perfor + adhere +
##     differ + extent + surg + node4
##
##            Df    AIC
## - age       1 5813.7
## - perfor    1 5813.9
## - sex       1 5814.2
## - adhere    1 5814.6
## <none>        5815.3
## - obstruct  1 5816.2
## - surg      1 5818.6
## - differ    2 5818.9
## - extent    3 5824.5
## - rx        2 5834.7
## - node4     1 5880.1
##
## Step:  AIC=5813.68
## Surv(time, status) ~ rx + sex + obstruct + perfor + adhere +
##     differ + extent + surg + node4
##
##            Df    AIC
## - perfor    1 5812.4
## - sex       1 5812.6
## - adhere    1 5812.9
```

```
## <none>        5813.7
## - obstruct  1 5814.8
## + age       1 5815.3
## - surg      1 5817.0
## - differ    2 5817.1
## - extent    3 5823.1
## - rx        2 5833.1
## - node4     1 5879.7
##
## Step:  AIC=5812.38
## Surv(time, status) ~ rx + sex + obstruct + adhere + differ +
##     extent + surg + node4
##
##            Df    AIC
## - sex       1 5811.2
## - adhere    1 5812.0
## <none>        5812.4
## + perfor    1 5813.7
## - obstruct  1 5813.9
## + age       1 5813.9
## - surg      1 5815.7
## - differ    2 5816.0
## - extent    3 5822.2
## - rx        2 5831.7
## - node4     1 5878.0
##
## Step:  AIC=5811.2
## Surv(time, status) ~ rx + obstruct + adhere + differ + extent +
##     surg + node4
##
##            Df    AIC
## - adhere    1 5810.9
## <none>        5811.2
## + sex       1 5812.4
## + perfor    1 5812.6
## + age       1 5812.8
## - obstruct  1 5812.9
## - surg      1 5814.4
## - differ    2 5814.7
## - extent    3 5821.1
## - rx        2 5830.1
## - node4     1 5877.5
##
## Step:  AIC=5810.87
## Surv(time, status) ~ rx + obstruct + differ + extent + surg +
##     node4
##
##            Df    AIC
## <none>        5810.9
## + adhere    1 5811.2
## + perfor    1 5811.9
## + sex       1 5812.0
## - obstruct  1 5812.5
## + age       1 5812.5
```

```
## - surg      1 5814.3
## - differ    2 5815.0
## - extent    3 5822.1
## - rx        2 5829.9
## - node4     1 5876.6
```

```
anova(r.model.aic)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##           loglik   Chisq Df Pr(>|Chi|)
## NULL     -2962.9
## rx       -2951.5 22.8126  2  1.113e-05 ***
## obstruct -2949.7  3.6738  1  0.0552739 .
## differ   -2942.7 13.9836  2  0.0009194 ***
## extent   -2931.6 22.2720  3  5.726e-05 ***
## surg     -2929.3  4.5254  1  0.0333961 *
## node4    -2895.4 67.7628  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise selection method selected model: Surv(time, status) ~ differ + rx + extent + surg + node4

4.2 Model Diagnostic

4.2.1 Check proportionality of hazard ratios

Log of Negative Log of Estimated Survival Function

```
r.differ.fit <- survfit(Surv(time, status) ~ differ, data = colon.recur)
cll.differ = ggsurvplot(r.differ.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.siz
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (Days)",
           legend.lab = c("1-well","2-moderate","3-poor"),
           legend.title = 'Differ',
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Recurrence by
```

```
r.extent.fit <- survfit(Surv(time, status) ~ extent, data = colon.recur)
cll.extent = ggsurvplot(r.extent.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.siz
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (Days)",
           legend.lab = c("1-submucosa","2-muscle","3-serosa", "4-contiguous structures"),
           legend.title = 'Extent',
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Recurrence by
```

```
r.surg.fit <- survfit(Surv(time, status) ~ surg, data = colon.recur)
cll.surg = ggsurvplot(r.surg.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (Days)",
           legend.lab = c("0-short","1-long"),
           legend.title = 'Surg',
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Recurrence by
```

```
r.node4.fit <- survfit(Surv(time, status) ~ node4, data = colon.recur)
cll.node4 = ggsurvplot(r.node4.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size =
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (Days)",
           legend.lab = c("0 = No", "1 =Yes"),
           legend.title = 'node4',
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Recurrence by
```

```
splots <- list()
splots[[1]] <- cll.differ
splots[[2]] <- cll.extent
splots[[3]] <- cll.surg
splots[[4]] <- cll.node4

cloglog_plot = arrange_ggsurvplots(splots, print = FALSE, ncol = 2, nrow = 2)
ggsave(cloglog_plot,file = "./plot/r.C-log-log-plots.pdf",width = 12,height = 15)
ggsave(cloglog_plot,file = "./plot/r.C-log-log-plots.png",width = 12,height = 15)
```

Schoenfeld residuals

```
r.residual = cox.zph(coxph(Surv(time, status) ~ differ + rx + extent + surg + node4, data = colon.recur)
r.residual
```

```
##          chisq df       p
## differ 19.074  2 7.2e-05
## rx      0.133  2 0.93568
## extent  1.716  3 0.63330
## surg    1.109  1 0.29236
## node4  11.088  1 0.00087
## GLOBAL 31.671  9 0.00023
```

```
residual_plot = ggcoxzph(r.residual, font.main = 10, font.x = 10, font.y = 10, font.tickslab = 8,
         point.alpha = 0.5, point.col = "grey25")

ggsave("./plot/d.schoenfeld residual_plots.pdf", arrangeGrob(grobs = residual_plot))
```

```
## Saving 6.5 x 4.5 in image
```

```
ggsave("./plot/d.schoenfeld residual_plots.png", arrangeGrob(grobs = residual_plot))
```

```
## Saving 6.5 x 4.5 in image
```

differ and node4 violates ph assumption

4.3 Modification for Violation of PH Assumption

```
#add interaction of covariate with function of time
colon.recur1 = colon.recur %>%
               mutate(differ_time = as.numeric(differ)*log(time),
                      node4_time = as.numeric(node4)*log(time))

r.model.inter = coxph(Surv(time, status) ~ rx + extent + surg + differ + node4 + differ_time + node4_ti
anova(r.model.inter)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##              loglik     Chisq Df Pr(>|Chi|)
## NULL        -2962.9
## rx          -2951.5   22.8126  2  1.113e-05 ***
## extent      -2938.6   25.9381  3  9.826e-06 ***
## surg        -2936.0    5.1632  1   0.023071 *
## differ      -2930.3   11.3183  2   0.003486 **
## node4       -2897.3   66.1219  1  4.239e-16 ***
## differ_time -1738.4 2317.7471  1  < 2.2e-16 ***
## node4_time  -1684.9  106.9277  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 Final model

```
r.final.model = coxph(Surv(time, status) ~ rx + extent + surg + differ + node4 + differ_time + node4_ti
summary(r.final.model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + extent + surg + differ +
##     node4 + differ_time + node4_time, data = colon.recur1)
##
##   n= 906, number of events= 458
##
##                   coef  exp(coef)   se(coef)       z Pr(>|z|)
## rx2          3.433e-02  1.035e+00  1.123e-01   0.306   0.7599
## rx3         -2.161e-01  8.057e-01  1.261e-01  -1.713   0.0866 .
## extent2      2.082e-01  1.231e+00  4.903e-01   0.425   0.6711
## extent3      2.086e-01  1.232e+00  4.622e-01   0.451   0.6518
## extent4      2.568e-01  1.293e+00  5.089e-01   0.505   0.6138
```

```
## surg1       -2.155e-02  9.787e-01  1.057e-01  -0.204   0.8385
## differ2      2.048e+01  7.853e+08  1.292e+00  15.855   <2e-16 ***
## differ3      3.622e+01  5.391e+15  2.236e+00  16.200   <2e-16 ***
## node41       1.066e+01  4.279e+04  1.247e+00   8.552   <2e-16 ***
## differ_time -2.960e+00  5.182e-02  1.800e-01 -16.448   <2e-16 ***
## node4_time  -1.801e+00  1.652e-01  2.130e-01  -8.456   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## rx2         1.035e+00  9.663e-01 8.304e-01 1.290e+00
## rx3         8.057e-01  1.241e+00 6.292e-01 1.032e+00
## extent2     1.231e+00  8.121e-01 4.711e-01 3.219e+00
## extent3     1.232e+00  8.118e-01 4.980e-01 3.048e+00
## extent4     1.293e+00  7.735e-01 4.768e-01 3.505e+00
## surg1       9.787e-01  1.022e+00 7.955e-01 1.204e+00
## differ2     7.853e+08  1.273e-09 6.244e+07 9.877e+09
## differ3     5.391e+15  1.855e-16 6.736e+13 4.314e+17
## node41      4.279e+04  2.337e-05 3.715e+03 4.929e+05
## differ_time 5.182e-02  1.930e+01 3.641e-02 7.373e-02
## node4_time  1.652e-01  6.055e+00 1.088e-01 2.507e-01
##
## Concordance= 0.972  (se = 0.003 )
## Likelihood ratio test= 2556  on 11 df,   p=<2e-16
## Wald test            = 685.8  on 11 df,   p=<2e-16
## Score (logrank) test = 2182  on 11 df,   p=<2e-16
```