Final Project

Zifeng(Robin) Zhan

2025-04-29

Inspect Raw Data

```
# Load data set
colon <- survival::colon
## quick structure
str(colon, give.attr = FALSE)
```

```
## 'data.frame':    1858 obs. of  16 variables:
##  $ id      : num  1 1 2 2 3 3 4 4 5 5 ...
##  $ study   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ rx      : Factor w/ 3 levels "Obs","Lev","Lev+5FU": 3 3 3 3 1 1 3 3 1 1 ...
##  $ sex     : num  1 1 1 1 0 0 0 0 1 1 ...
##  $ age     : num  43 43 63 63 71 71 66 66 69 69 ...
##  $ obstruct: num  0 0 0 0 0 0 1 1 0 0 ...
##  $ perfor  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adhere  : num  0 0 0 0 1 1 0 0 0 0 ...
##  $ nodes   : num  5 5 1 1 7 7 6 6 22 22 ...
##  $ status  : num  1 1 0 0 1 1 1 1 1 1 ...
##  $ differ  : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ extent  : num  3 3 3 3 2 2 3 3 3 3 ...
##  $ surg    : num  0 0 0 0 0 0 1 1 1 1 ...
##  $ node4   : num  1 1 0 0 1 1 1 1 1 1 ...
##  $ time    : num  1521 968 3087 3087 963 ...
##  $ etype   : num  2 1 2 1 2 1 2 1 2 1 ...
```

```
## how much is actually missing?
na_totals <- sapply(colon, \(x) sum(is.na(x)))
na_totals
```

```
##       id    study       rx      sex      age obstruct   perfor   adhere
##        0        0        0        0        0        0        0        0
##    nodes   status   differ   extent     surg    node4     time    etype
##       36        0       46        0        0        0        0        0
```

```
round(na_totals[na_totals > 0] / nrow(colon) * 100, 2)
```

```
##  nodes differ
##   1.94   2.48
```

```
# nodes has 1.94% missing value and differ has 2.48% missing values
# missingness is small, delete the obs
```

```
# Delete rows with NA
colon <- colon[complete.cases(colon), ]

# Convert integer to factor
colon_clean <- colon %>%
  mutate(sex = factor(sex, labels = c("Female", "Male")),
         obstruct = as.factor(obstruct),
         perfor = as.factor(perfor),
         adhere = as.factor(adhere),
         differ = factor(differ, 1:3, labels = c("well", "moderate", "poor")),
         node4 = as.factor(node4),
         surg = factor(surg, 0:1, labels = c("short", "long")),
         extent = as.factor(extent),
         etype = as.factor(etype))
```

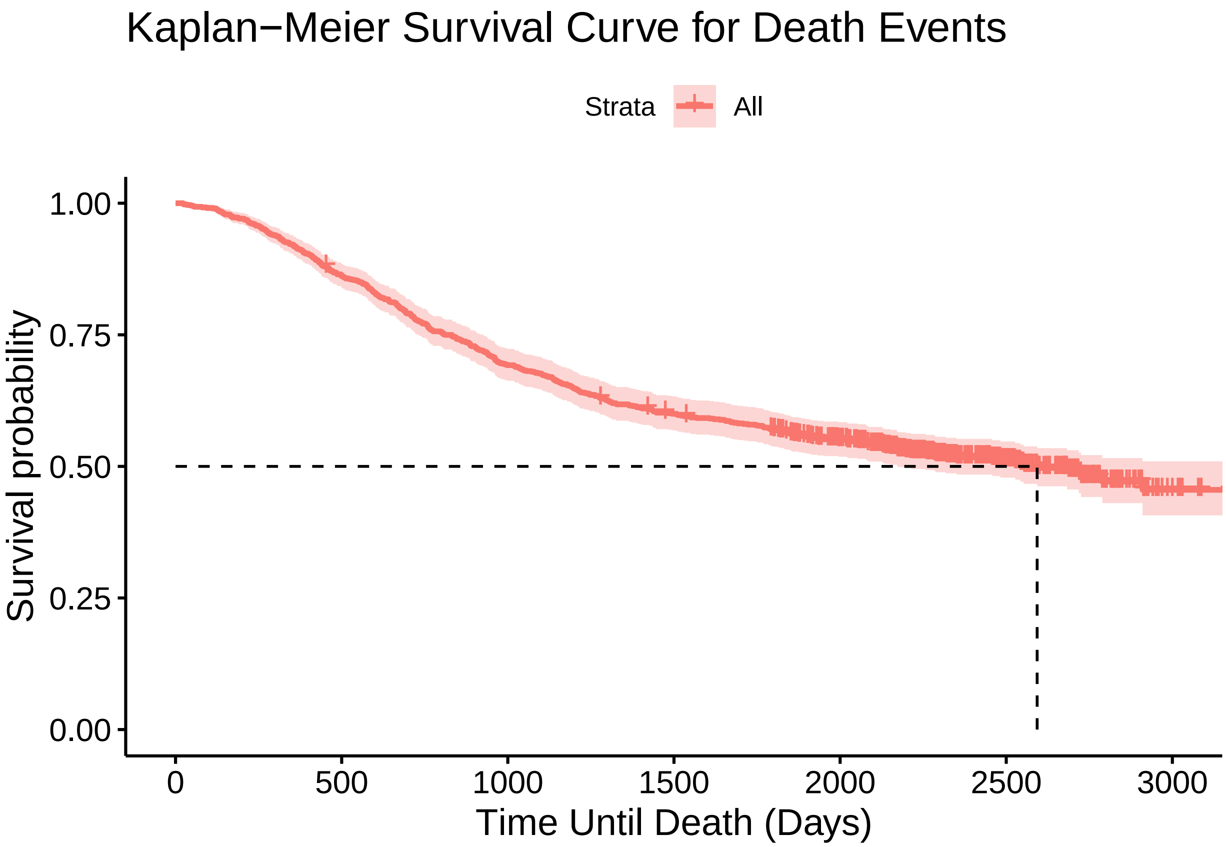
Death Event

```
# Data set for death event
# Filter the data for death event
colon_death <- colon_clean[colon_clean$etype == 2, ]
## ensure one row for each id
stopifnot(all(!duplicated(colon_death$id)))
```

KP

```
# Fit the Kaplan-Meier model for the death event
km_fit_death <- survfit(Surv(time, status) ~ 1, data = colon_death)

# Plot the Kaplan-Meier curve for death events
ggsurvplot(km_fit_death, data = colon_death,
           title = "Kaplan-Meier Survival Curve for Death Events",
           xlab = "Time Until Death (Days)",
           surv.median.line = 'hv',
           break.time.by = 500)
```



```r
# Median survival time
Death_med <- surv_median(km_fit_death)
print(Death_med)
```

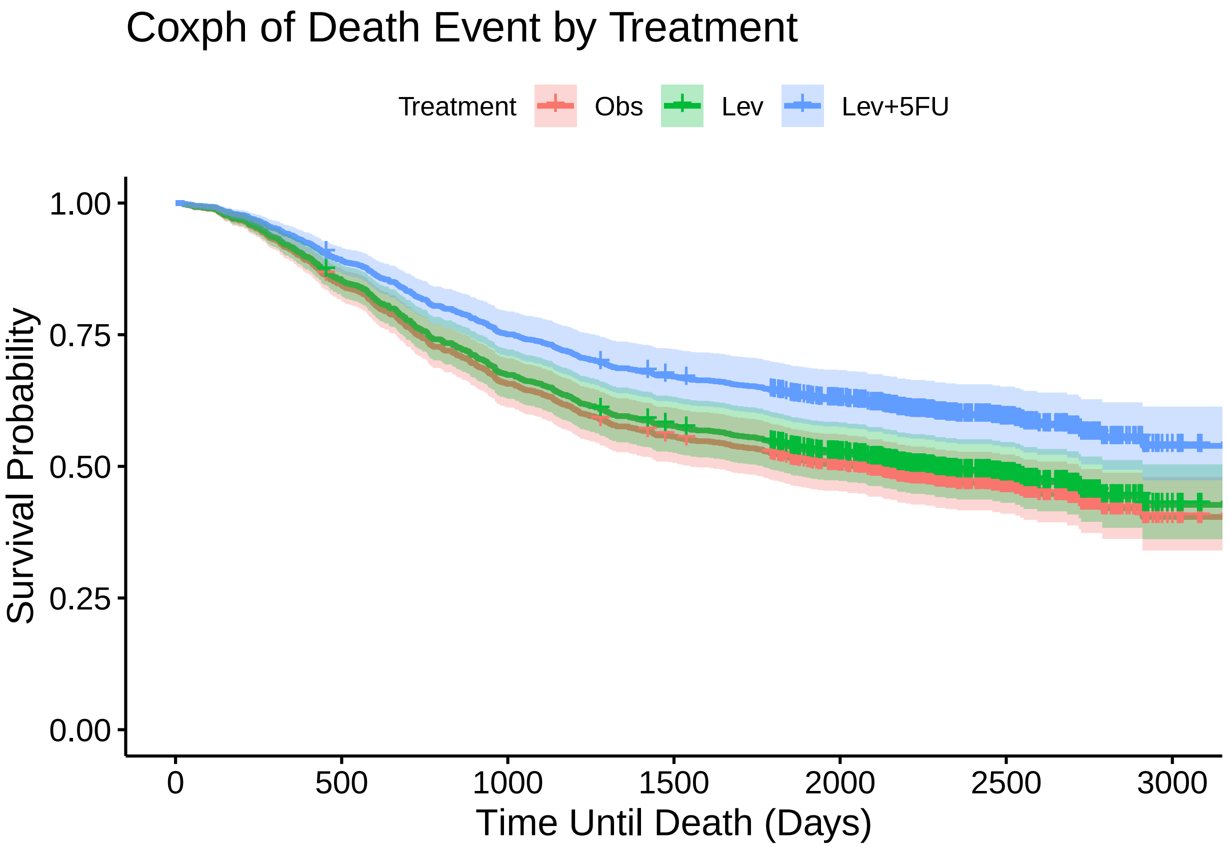
```
##   strata median lower upper
## 1    All   2593  2174    NA
```

The Kaplan-Meier curve declines slowly and almost linearly over the 3000 days follow-up and the median survival time is 2593 days. At the beginning of the study before one year (365 days), the survival probability is roughly above 90%, which indicate that patients in the study begin with a near-perfect chance of remaining alive. Also, the numerous tick marks in the late tail indicate that many individuals were censored alive at the later stage of the study, which is common because it is hard to follow-up for a long period.

Cox

```r
# Coxph model with rx(treatment) as covariate
cox_death <- coxph(Surv(time, status) ~ rx, data = colon_death)
# Create fit for different treatment
fit_death <- survfit(cox_death, newdata = data.frame(rx = c("Obs", "Lev", "Lev+5FU")))

# Plot fit for coxph
ggsurvplot(fit_death, data = colon_death, conf.int = TRUE,
           ylab = "Survival Probability",
           xlab = "Time Until Death (Days)",
           title = "Coxph of Death Event by Treatment",
           legend.title = "Treatment",
           legend.lab = levels(colon_death$rx),
           break.time.by = 500)
```



```
# median survival time
cox_med <- surv_median(fit_death)
print(cox_med)
```

```
##   strata median lower upper
## 1      1   2052  1550  2718
## 2      2   2257  1767    NA
## 3      3     NA  2789    NA
```

```
# Summary of PH model
summary(cox_death)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon_death)
##
##   n= 888, number of events= 430
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.06269   0.93923  0.11319 -0.554  0.57967
## rxLev+5FU -0.38280   0.68195  0.12110 -3.161  0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9392      1.065    0.7524    1.1725
## rxLev+5FU    0.6819      1.466    0.5379    0.8646
##
## Concordance= 0.537  (se = 0.013 )
## Likelihood ratio test= 11.41  on 2 df,   p=0.003
## Wald test            = 10.9  on 2 df,   p=0.004
## Score (logrank) test = 11.02  on 2 df,   p=0.004
```

This plot show the survival curves for three treatment after fitting a Cox model with only treatment (`rx`) as the covariate. `Lev+5FU` (blue) curve locate above the other two curves, which might indicate that Levamisole+5-FU (`Lev+5FU`) can increase patients' survival rate. And the its survival probability decrease from 1 and end above 0.5, show that most of patient survive after the study.

`Lev` (green) and `obs` (red) lines do not show much difference. The median survival time for `obs` is 2052 days and for `Lev` is 2257 days greater than `obs`. 95% confidence interval of median survival time for `obs` is (1550, 2718) and the lower bound for confidence interval for `Lev` is 1767. Since the two confidence interval is overlap, there is not statistically significant different between the median survival time of `obs` and `Lev`.

Since the p-value of likelihood ratio test is 0.003 less than $\alpha = 0.05$, there is sufficient evidence to conclude that `rx` has significant impact on the survival time of patients.

Since hazard ratio for `Lev` is 0.9392, patients on `Lev` has 6.08% lower hazard rate than observation. Also the 95% confidence interval for `Lev` is (0.7524, 1.1725) including 1. Thus, Levamisole does not have significant impact on the survival probability.

The hazard ratio for `Lev+5FU` is 0.6819, `Lev+5FU` has 32% lower hazard rate than `Observation`. Also, 95 confidence interval (0.5379, 0.8646) does not include 1. Treatment `Lev+5FU` significantly lower the hazard rate and increase the survival probability of patient.

AIC

```r
# 1st Covariate
# List of Covariate to test
uni_vars <- c("obstruct", "adhere", "nodes", "node4", "differ",
              "extent", "surg", "perfor", "age", "sex")

## 2.  Build one model per variable
uni_models <- map(uni_vars, \(v)
  coxph(as.formula(paste("Surv(time, status) ~ rx + ", v)),
        data = colon_death)
) |> set_names(uni_vars)

## 3.  Grab AIC
aic_tbl <- map_dbl(uni_models, AIC) |>
           sort() |>
           round(2)

aic_tbl
```

```
##    node4    nodes   extent   differ   adhere obstruct     surg      age
##  5446.81  5460.49  5507.84  5519.59  5525.62  5526.58  5527.03  5529.74
##   perfor      sex
##  5530.06  5530.38
```

```
# 1st = node4
```

```
# 2nd Covariate
uni_vars2 <- c("obstruct", "adhere", "differ",
              "extent", "surg", "perfor", "age", "sex")

uni_models2 <- map(uni_vars2, \(v)
  coxph(as.formula(paste("Surv(time, status) ~ rx + node4 + ", v)),
        data = colon_death)
) |> set_names(uni_vars2)

aic_tbl2 <- map_dbl(uni_models2, AIC) |>
           sort() |>
           round(2)

aic_tbl2
```

```
##   extent   differ obstruct   adhere     surg      age   perfor      sex
##  5432.64  5443.44  5443.53  5444.36  5444.46  5445.52  5448.34  5448.79
```

```
# 2nd = extent
```

```
# 3rd Covariate
uni_vars3 <- c("obstruct", "adhere", "differ",
               "surg", "perfor", "age", "sex")

uni_models3 <- map(uni_vars3, \(v)
  coxph(as.formula(paste("Surv(time, status) ~ rx + node4 + extent + ", v)),
        data = colon_death)
) |> set_names(uni_vars3)

aic_tbl3 <- map_dbl(uni_models3, AIC) |>
           sort() |>
           round(2)

aic_tbl3
```

```
##     surg   differ obstruct      age   adhere   perfor      sex
##  5429.86  5430.15  5431.04  5431.17  5431.66  5434.44  5434.61
```

```
# 3rd  = surg
```

```
# 4th Covariate
uni_vars4 <- c("obstruct", "adhere", "differ",
               "perfor", "age", "sex")

uni_models4 <- map(uni_vars4, \(v)
  coxph(as.formula(paste("Surv(time, status) ~ rx + node4 + extent + surg + ", v)),
```

```r
        data = colon_death)
) |> set_names(uni_vars4)

aic_tbl4 <- map_dbl(uni_models4, AIC) |>
           sort() |>
           round(2)

aic_tbl4
```

```
##   differ obstruct      age   adhere   perfor      sex
##  5427.51  5428.30  5428.62  5429.07  5431.68  5431.85
```

```r
# 4th = differ
```

```r
# 5th Covariate
uni_vars5 <- c("obstruct", "adhere",
               "perfor", "age", "sex")

uni_models5 <- map(uni_vars5, \(v)
  coxph(as.formula(paste("Surv(time, status) ~ rx + node4 + extent + surg +
                          differ + ", v)),
        data = colon_death)
) |> set_names(uni_vars5)

aic_tbl5 <- map_dbl(uni_models5, AIC) |>
           sort() |>
           round(2)

aic_tbl5
```

```
## obstruct      age   adhere   perfor      sex
##  5425.77  5426.61  5427.38  5429.36  5429.50
```

We selected extra covariates by forward AIC while always keeping treatment (`rx`) in the model. Adding `node4`, `extent`, and `surg` each cut AIC by > 2 points, and `differ` lowered it by another 2.4; `obstruct` reduced AIC by < 2. Because 2 points is the standard threshold for a meaningful gain, we stopped at `rx + node4 + extent + surg + differ`. This captures nearly all improvement in fit without adding unnecessary parameters.

Full Model

```r
full_death <- coxph(Surv(time, status) ~ node4 + extent +
                      surg + differ + rx, data = colon_death)
summary(full_death)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ node4 + extent + surg +
##     differ + rx, data = colon_death)
##
```

```
##   n= 888, number of events= 430
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## node41          0.90682   2.47645  0.10076  9.000  < 2e-16 ***
## extent2         0.58647   1.79764  0.60331  0.972  0.33100
## extent3         1.10438   3.01735  0.58214  1.897  0.05781 .
## extent4         1.56152   4.76605  0.61527  2.538  0.01115 *
## surglong        0.23256   1.26182  0.10625  2.189  0.02862 *
## differmoderate -0.08838   0.91541  0.16812 -0.526  0.59910
## differpoor      0.23602   1.26620  0.19489  1.211  0.22587
## rxLev          -0.04548   0.95554  0.11429 -0.398  0.69066
## rxLev+5FU      -0.37257   0.68896  0.12185 -3.058  0.00223 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## node41            2.4764     0.4038    2.0326    3.0172
## extent2           1.7976     0.5563    0.5510    5.8646
## extent3           3.0173     0.3314    0.9641    9.4437
## extent4           4.7661     0.2098    1.4271   15.9176
## surglong          1.2618     0.7925    1.0246    1.5540
## differmoderate    0.9154     1.0924    0.6584    1.2727
## differpoor        1.2662     0.7898    0.8642    1.8552
## rxLev             0.9555     1.0465    0.7638    1.1954
## rxLev+5FU         0.6890     1.4515    0.5426    0.8748
##
## Concordance= 0.662  (se = 0.013 )
## Likelihood ratio test= 126.3  on 9 df,   p=<2e-16
## Wald test            = 129.1  on 9 df,   p=<2e-16
## Score (logrank) test = 139.3  on 9 df,   p=<2e-16
```

After adjusting for the four strongest prognostic factors—`node4`, `extent`, `surg`, and `differ`—the overall likelihood-ratio test is highly significant ($p < 2Ö10^{-16}$), confirming that the set of covariates is statistically significant to explain variation in survival model. From the summary of the cox proptional model, we can observe the following effect of treatment and prognostic covariate:

Treatment effect:
The combination therapy `Levamisole+5-FU` has statistically significant survival benefit, reducing the hazard of death by approximately 31% with (HR = 0.689, 95% CI 0.54˘0.88). `Levamisole` alone does not show significant benefit because the 95% CI ($0.7638 - 1.1954$) include 1.

Prognostic covariates:
`node4`: having more than 4 positive lymph nodes has hazard ratio of 2.4764 and significantly increase the hazard risk by 147% compared to less than 4 lymph nodes (95% CI 2.03˘3.02).
`extent`: Contiguous structures of local spread (`extent` = 4) raises the hazard by 377% compared to to submucosa of local spread (`extent`=1) (HR = 4.77, 95 % CI 1.96–15.9).
`surg`: Long time from surgery to registration (`surg` = 1) also raise the hazard rate by 26.18% compared to shorter time (`surg`= 0) (HR = 1.26, 95% CI 1.02˘1.55).