Colon Cancer Survival Analysis

Calder Glass, Kotaro Ito, Robin Zhan

2025-06-01

Introduction

We explore the *colon* dataset found in the R "survival" package, containing data observed from a clinical trial where stage B/C colon cancer patients receive adjuvant chemotherapy. 929 independent patients–484 men and 445 women–were randomly assigned between two treatments and a control group: Levamisole, Levamisole and Fluorouracil, and control(denoted as `Lev`, `Lev+5FU`, and `Obs` in the dataset). "Levamisole is a low-toxicity compound that was originally used to treat worm infestations in animals", while "5-FU is a moderately toxic chemotherapy agent" used to treat cancer.[1]

Patients were then observed until one of two events occurred: recurrence or death (denoted as "1" and "2" in its respective order under column `etype`). The time of occurrence, in days, was then recorded to later investigate and determine whether or not different treatments were effective in keeping the patients alive. Each patient in the dataset, identified by their `id`, has two rows for both recurrence and death. The status column indicates whether or not the event occurred or not ("0" indicates no and "1" indicates yes). If a patient has been recorded for 3000 days for both recurrence and death and the status remains 0 for both, it signifies that they did not experience any event for 3000 days and dropped out of the study for unknown reasons. Figure 1 and 2 below represents the Kaplan-Meier Survival Curve after splitting the dataset by `etype` (recurrence and death). The convex shape of Figure 1 conveys that many recurrences occur early on while the curve for death events show that deaths in patients are gradual and consistent.

Figure 1: KP for Recurrence Events

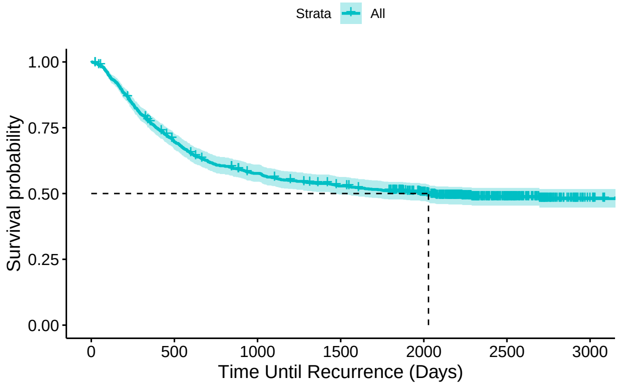
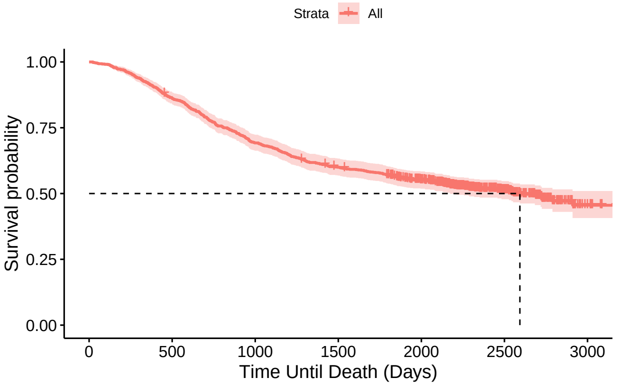


Figure 2: KP for Death Events



The dataset contains the `id`, `age`(in years), `sex` ("1" indicates man and "0" indicates woman), `rx` (the treatment type or control), `obstruct` ("1" indicates a colon obstructed by a tumor and "0" indicates no obstruction), `perfor` ("1" indicates a perforated colon and "0" indicates no perforation), `adhere` ("1" indicates cancer adhering to other organs and "0" indicates no adherence), `nodes` (the number of lymph nodes with colon cancer), `time` (time until event occurrence or censoring), `status` (whether or not the event occurred or

[1] colon: Chemotherapy for Stage B/C colon cancer. Colon Cancer. Retrieved May 1st, 2025, from https://rdrr.io/cran/survival/man/colon.html

not), `differ` ("3" indicates quickly growing cancer, "2" indicates moderate growth, and "1" indicates slowly growing and less likely to spread), `extent` (describes the spread of the tumor and ranges from 1-4, where "1" indicates that the tumor is limited to the inner lining of the colon and "4" indicates invasion of tumor to nearby organs and tissues), `surg` ("1" indicates a long time between initial surgery and registering to the study while "0" indicates a short time), `node4` ("1" indicates a patient has more than four positive lymph nodes and "0" indicates four or less), and `etype` (recurrence or death event) of each patient.[1]

Taking all of the covariates we listed above into consideration, our aim is to determine whether or not a specific treatment has a significant effect on the survival of the patient. Our secondary objective is to assess which covariate(s) have a significant effect on the hazard risk. In the course of the analysis, we omit observations with N/A values, reducing our final dataset to 888 independent patients. On a deeper level, we are interested in how the effect of the treatment change after recurrence of colon cancer, and by how much relative to each other's hazard ratios. A five percent significance level (0.05) will be used to balance the risk of false positives to detect meaningful effects.

Model Fitting

With the clinical context established and the relevant covariates explained, we begin to evaluate the effects of treatment and other factors on the patient. Given that our dataset includes two types of events — recurrence of cancer and death — we begin by modeling these outcomes separately using marginal Cox proportional hazards models. This allows us to estimate the hazard associated with each covariate for each event type independently and by fitting separate Cox models for recurrence and death, we can assess whether specific treatments or patient characteristics are associated with an increased or decreased risk for each type of event.

Marginal Model: Recurrence

Kaplan-Meier Estimate for Recurrence

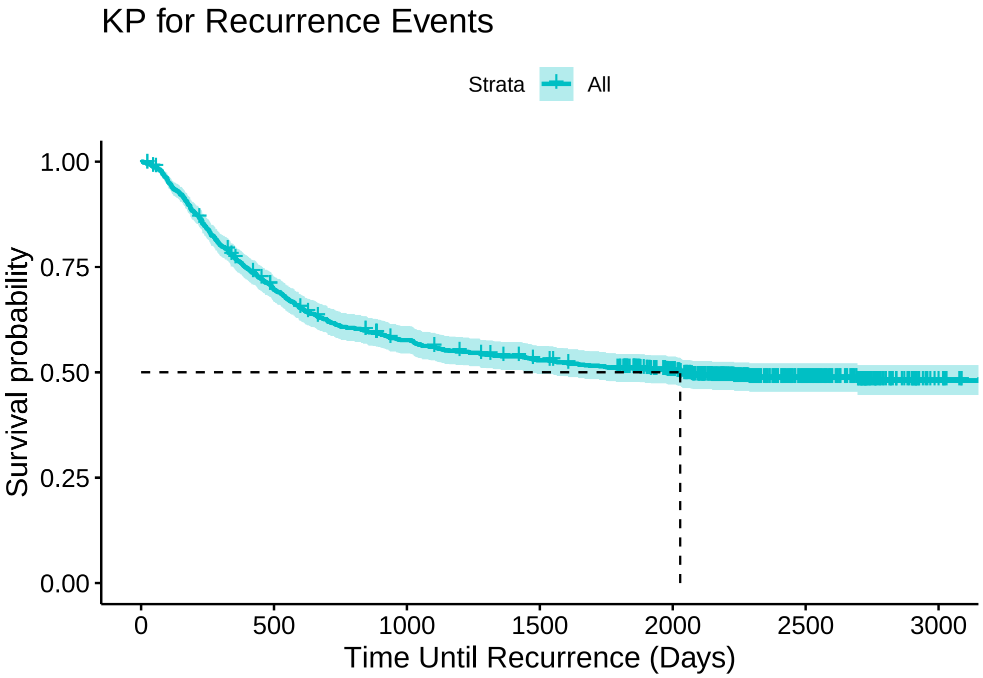


Table 1: KP Median Survival Time of Recurrence

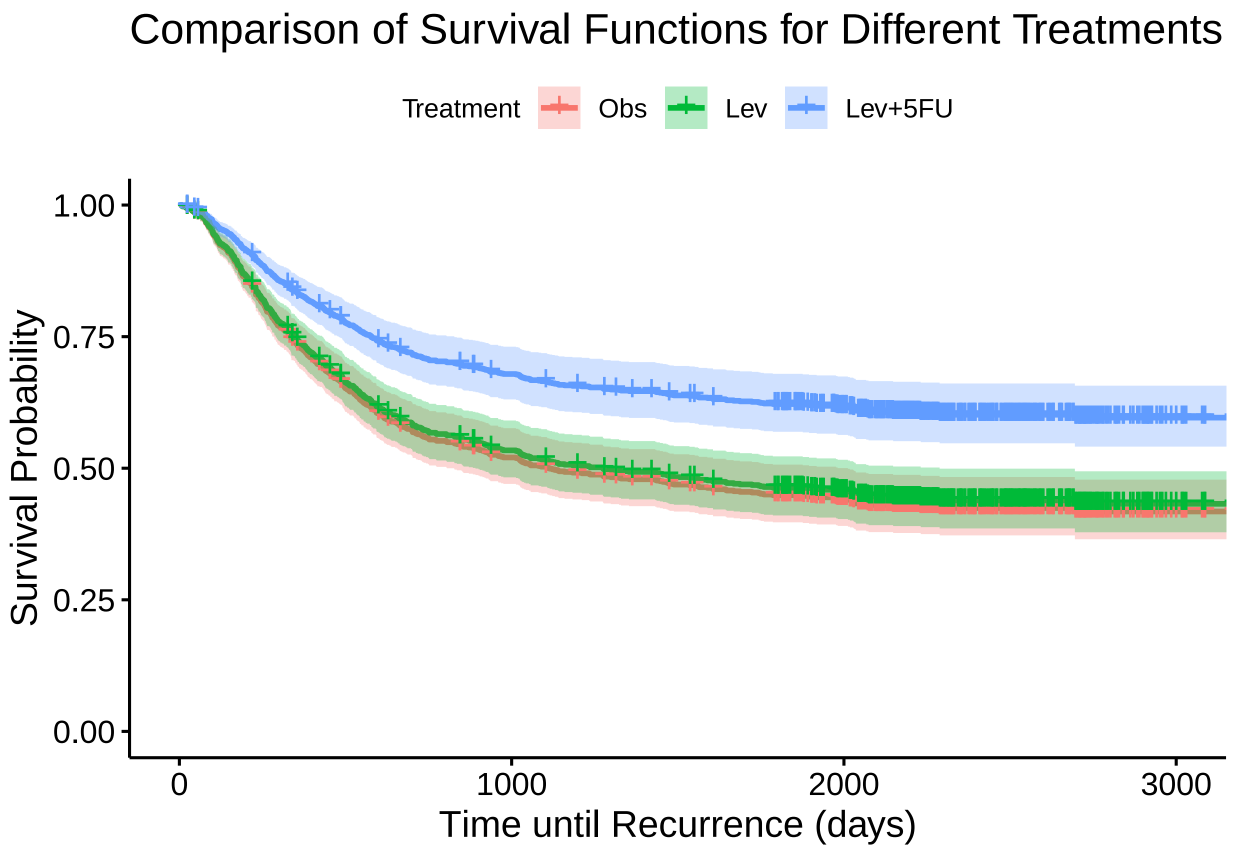
strata	median	lower	upper
All	2028	1471	NA

Most of the decay in the survival probability for the Kaplan-Meier curve occurs in the first 1500 days. The estimate hits the median survival probability at 2028 days.

The thick density of tick marks after approximately 1750 days suggest that many subjects were censored, either passing away after the study or dropping out before the study concluded.

Proportional Hazard Model for Recurrence

```
recurrence_obj = coxph(Surv(time, status) ~ rx, data = recurrence_data)
recurrence_fit = survfit(recurrence_obj,
                         newdata = data.frame(rx= c("Obs", "Lev", "Lev+5FU")))
ggsurvplot(recurrence_fit, data = recurrence_data,
          conf.int = TRUE,
           title = "Comparison of Survival Functions for Different Treatments",
           xlab = "Time until Recurrence (days)",
           ylab = "Survival Probability",
           legend.title = "Treatment",
           legend.lab = levels(recurrence_data$rx))
```



```r
# Median Survival Time for each Treatment
recurrence_cox_median = surv_median(recurrence_fit)
recurrence_cox_median$strata <- c("Obs", "Lev", "Lev+5FU")
knitr::kable(recurrence_cox_median, caption = "Coxph Median Survival Time of Recurrence by Treatment")
```

Table 2: Coxph Median Survival Time of Recurrence by Treatment

strata	median	lower	upper
Obs	1114	805	2012
Lev	1275	891	2288
Lev+5FU	NA	NA	NA

From the inital plot, the survival curve for Levamisole & Fluoracil appears to be greater than the other treatments, so the survival probabilities for the patients on Levamisole + Fluoracil are higher than those on the other treatments.

This would suggest that Levamisole & Fluoracil could improve survival probabilities for patients.

The other two survival curves, only Levamisole and observed/control, have very similar survival curves with huge overlaps between their confidence intervals. This would suggest that only taking Levamisole does not have a particular significant boost on the survival probability of colon cancer for its patients. The strong overlap is also reflected in the median survival times, which are less than 200 days apart for only Levamisole and the control group. The group taking both Levamisole and Fluoracil does not have a recorded median survival time because its survival curve doesn't decrease to the 50% mark within the allocated time of the study.

```r
summary(recurrence_obj)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = recurrence_data)
##
##   n= 888, number of events= 446
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev      -0.03996   0.96083  0.10967 -0.364    0.716
## rxLev+5FU  -0.52289   0.59280  0.12118 -4.315  1.6e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9608      1.041    0.7750    1.1912
## rxLev+5FU    0.5928      1.687    0.4675    0.7517
##
## Concordance= 0.556  (se = 0.013 )
## Likelihood ratio test= 23.15  on 2 df,   p=9e-06
## Wald test            = 21.5  on 2 df,   p=2e-05
## Score (logrank) test = 21.95  on 2 df,   p=2e-05
```

The 95% confidence interval for the hazard ratio of Levamisole + Fluoracil is $(0.4675, 0.7517)$ and the hazard ratio itself is 0.5928, only further suggesting that the combined treatment has a significant role in increasing survival probability to colon cancer for its patients. Meanwhile, the 95% confidence interval for the hazard ratio only Levamisole is $(0.775, 1.1912)$ and the hazard ratio itself is 0.9608, so there is very little difference between taking or not taking only Levamisole treatment in terms of the hazard rate.

Model Selection Using AIC

To identify a well-fitting but not over-fitting set of covariates for each Cox model, we applied forward stepwise selection driven by Akaike's Information Criterion (AIC). Beginning with a base model that always included the treatment variable (`rx`), candidate prognostic factors were added one at a time. At each step we chose the covariate that produced the lowest AIC for the whole model or the greatest drop in AIC compared to previous model.

For the AIC tests, the covariates `study` and `id` were not included. The `id` covariate is the same as the observation number, it does not have contextual significance to the event of relapse or death from colon cancer. The `study` covariate is not included as all of the subjects are from the same study. Then the list of covariates put into the models and compute its AIC would be `obstruct`, `adhere`, `nodes`, `node4`, `differ`, `extent`, `surg`, `perfor`, `sex`, `age`.

```r
# Level 1:
# Construct a list of covariates to put into the models:
recurrence_covariates = c("obstruct", "adhere", "nodes", "node4", "differ",
                          "extent", "surg", "perfor", "sex", "age")
# building a model per covariate by pasting the given covariate into the formula
# the set_names function helps to clear up which AIC value corresponds to
# which model when performing the AIC function

recurrence_models = map(recurrence_covariates, \(v)
                        coxph(as.formula(paste("Surv(time, status) ~ rx + ", v)),
                              data = recurrence_data)) |>
  set_names(recurrence_covariates)

aic_lvl1 = map_dbl(recurrence_models, AIC) |>
  sort()

# aic_lvl1
# 1st = node4
recurr1.1 <- coxph(Surv(time, status) ~ rx + node4, data = recurrence_data)

# Level 2:
# Construct a list of covariates to put into the models:
recurrence_covariates2 = c("obstruct", "adhere", "differ", "extent", "surg",
                           "perfor", "sex", "age")

recurrence_models2 = map(recurrence_covariates2, \(v)
                         coxph(as.formula(paste("Surv(time, status) ~ rx +
                                                node4 + ", v)),
                               data = recurrence_data)) |>
  set_names(recurrence_covariates2)

aic_lvl2 = map_dbl(recurrence_models2, AIC) |>
  sort()
# 2nd extent
recurr1.2 <- coxph(Surv(time, status) ~ rx + node4 + extent
                   , data = recurrence_data)

# aic_lvl2

# Level 3:
# Construct a list of covariates to put into the models:
```

```
recurrence_covariates3 = c("obstruct", "adhere", "differ", "surg", "perfor",
                           "sex", "age")

recurrence_models3 = map(recurrence_covariates3, \(v)
                         coxph(as.formula(paste(
                           "Surv(time, status) ~ rx + node4 + extent + ", v)),
                           data = recurrence_data)) |>
  set_names(recurrence_covariates3)

aic_lvl3 = map_dbl(recurrence_models3, AIC) |>
  sort()
# aic_lvl3
# 3rd = surg
recurr1.3 <- coxph(Surv(time, status) ~ rx + node4 + extent + surg
                   , data = recurrence_data)

# Level 4:
# list of covariates to put into the models
recurrence_covariates4 = c("obstruct", "adhere", "differ", "perfor", "sex",
                           "age")
recurrence_models4 = map(recurrence_covariates4, \(v)
                         coxph(as.formula(paste("Surv(time, status) ~ rx +
                                                node4 + extent + surg + ", v)),
                               data = recurrence_data)) |>
  set_names(recurrence_covariates4)

aic_lvl4 = map_dbl(recurrence_models4, AIC) |>
  sort()
# aic_lvl4
# 4th = differ
recurr1.4 <- coxph(Surv(time, status) ~ rx + node4 + extent + surg + differ,
                   data = recurrence_data)

# Level 5:
# list of covariates to put into the models
recurrence_covariates5 = c("adhere", "obstruct", "perfor", "sex", "age")
recurrence_models5 = map(recurrence_covariates5, \(v)
                         coxph(as.formula(paste("Surv(time, status) ~ rx +
                                                node4 + extent + surg +
                                                differ + ", v)),
                               data = recurrence_data)) |>
  set_names(recurrence_covariates5)

aic_lvl5 = map_dbl(recurrence_models5, AIC) |>
  sort()
# aic_lvl5
# 5th = obstruct
recurr1.5 <- coxph(Surv(time, status) ~ rx + node4 + extent + surg + differ
                   + obstruct,
                   data = recurrence_data)

# Level 6:-----------------------------------
# list of covariates to put into the models
```

```
recurrence_covariates6 = c("adhere", "perfor", "sex", "age")

# building a model per covariate by pasting the given covariate into the formula

# the set_names function helps to clear up which AIC value corresponds to which model when performing t

recurrence_models6 = map(recurrence_covariates6, \(v)
                         coxph(as.formula(paste("Surv(time, status) ~
                                                rx + node4 + extent + surg +
                                                differ + obstruct + ", v)),
                               data = recurrence_data)) |>
  set_names(recurrence_covariates6)

aic_lvl6 = map_dbl(recurrence_models6, AIC) |>
  sort()
# aic_lvl6
# 6th = Adhere
recurr1.6 <- coxph(Surv(time, status) ~ rx + node4 + extent + surg + differ
                   + obstruct + adhere,
                   data = recurrence_data)

recurAIC <- AIC(recurr1.1, recurr1.2, recurr1.3, recurr1.4, recurr1.5, recurr1.6)
rownames(recurAIC) <- c("rx+node4", "rx+node4+extent", "rx+node4+extent+surg",
                        "rx+node4+extent+surg+differ", "rx+node4+extent+surg+differ+obstruct", "rx+node
knitr::kable(recurAIC, caption = "Step AIC for Recurrence")
```

Table 3: Step AIC for Recurrence

	df	AIC
rx+node4	3	5666.763
rx+node4+extent	6	5650.922
rx+node4+extent+surg	7	5647.952
rx+node4+extent+surg+differ	9	5646.470
rx+node4+extent+surg+differ+obstruct	10	5645.417
rx+node4+extent+surg+differ+obstruct+adhere	11	5645.561

When choosing the first covariate `node4`, the binary variable for whether the patient had more than 4 positive lymph nodes, had the lowest AIC. Since the `nodes` covariate and `node4` covariate are closely related, the `nodes` covariate will be skipped for next iteration. Based on the logic of forward selection, the model stops adding covariate until checking covariate `adhere` because the AIC start increasing.

However, `obstruct` and `differ` only decreased the AIC by less than 2, which is the standard a significant AIC gain. `obstruct` did not significantly improve the fit of the model and adding an extra covariate might over-fitted the model. Therefore, the the final cox proportional hazard model for recurrence event only use `rx`, `surg`, `extent`, and `node4`.

Full Coxph Model for Recurrence

```
full_recurrence <- coxph(Surv(time, status) ~ rx + node4 + extent + surg, data = recurrence_data)
summary(full_recurrence)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + node4 + extent + surg,
##     data = recurrence_data)
##
##   n= 888, number of events= 446
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.02871   0.97170  0.11021 -0.260   0.7945
## rxLev+5FU -0.49312   0.61072  0.12169 -4.052 5.08e-05 ***
## node41     0.84049   2.31750  0.09908  8.483  < 2e-16 ***
## extent2    0.26290   1.30070  0.53001  0.496   0.6199
## extent3    0.84593   2.33014  0.50468  1.676   0.0937 .
## extent4    1.38116   3.97953  0.54157  2.550   0.0108 *
## surg1      0.23556   1.26562  0.10399  2.265   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9717     1.0291    0.7829    1.2060
## rxLev+5FU    0.6107     1.6374    0.4811    0.7752
## node41       2.3175     0.4315    1.9085    2.8142
## extent2      1.3007     0.7688    0.4603    3.6756
## extent3      2.3301     0.4292    0.8666    6.2656
## extent4      3.9795     0.2513    1.3767   11.5032
## surg1        1.2656     0.7901    1.0323    1.5517
##
## Concordance= 0.656  (se = 0.013 )
## Likelihood ratio test= 120.3  on 7 df,   p=<2e-16
## Wald test            = 122.2  on 7 df,   p=<2e-16
## Score (logrank) test = 129.6  on 7 df,   p=<2e-16
```

From the likelihood ratio test, the overall p-value is less than $2e - 16$, which is much less than the critical value/significance level of 0.05.

`rx`: The hazard rate for patients who took the treatment with just Levamisole is only 2.83% less hazardous than taking no treatment at all. Those who took Fluoracil in addition to Levamisole benefited with a hazard ratio of 0.6107, 38.93% less hazardous than no treatment at all.

`node4`: Patients who had more than 4 positive lymph nodes had over double the hazard rate of those who didn't.

`extent`: As the spread of the tumor developed from muscles to continguous structures, the hazard ratio to those who only had submucosa development increased to as high as 3.9795 times as likely to suffer a recurrence of colon cancer.

`surg`: Patients with a long time from their initial surgery to registration in the study had a 26.56% greater hazard rate than those with a shorter time interval.

Marginal Model: Death

Kaplan-Meier Estimate for Death

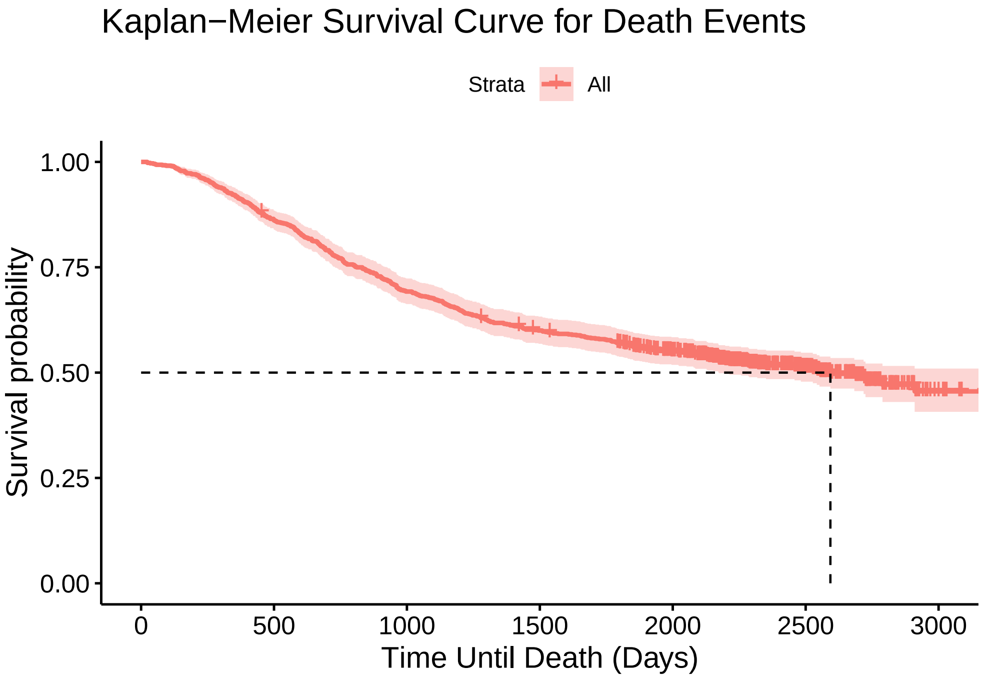


Table 4: KP Median Survival Time of Death

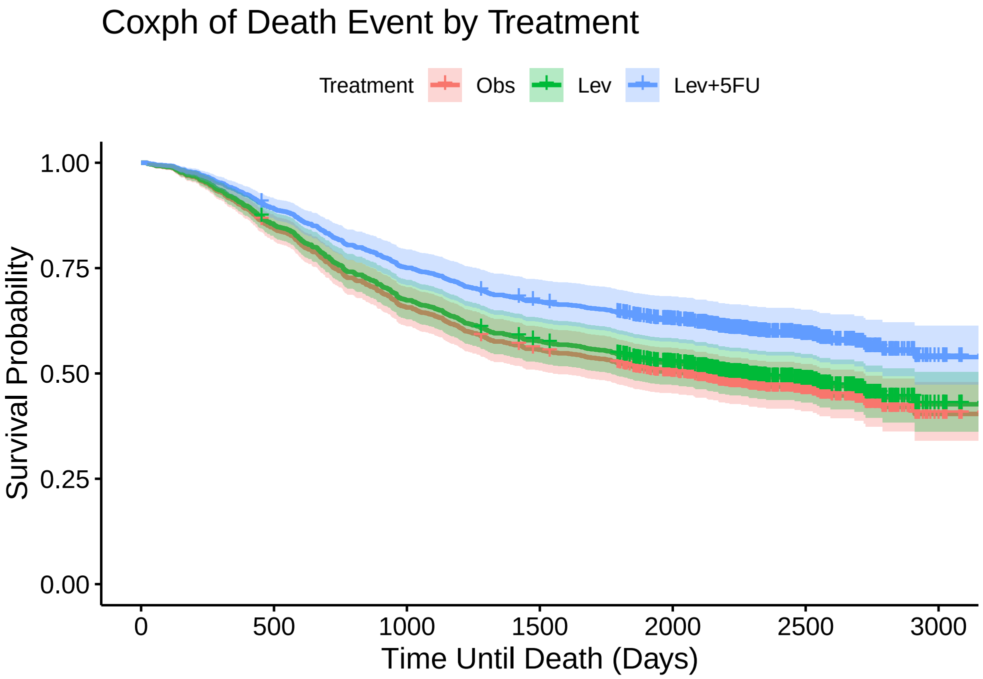
strata	median	lower	upper
All	2593	2174	NA

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.

Proportional Hazard Model for Death

```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)
cox_med$strata <- c("Obs", "Lev", "Lev+5FU")
knitr::kable(cox_med, caption = "Coxph Median Survival Time of Death by Treatment")
```

Table 5: Coxph Median Survival Time of Death by Treatment

strata	median	lower	upper
Obs	2052	1550	2718
Lev	2257	1767	NA
Lev+5FU	NA	2789	NA

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`.

```r
# 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
```

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.

Model Selection Using AIC

The cox proportional hazard model for death event applied forward step-wise selection driven by AIC and selected extra covariate based the on the same process in recurrence event. And the model tested the same list of covariate in the recurrence event.

```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) |>
```

```r
           sort() |>
           round(2)
# 1st = node4
# aic_tbl
death1.1 <- coxph(Surv(time, status) ~ rx+ node4, data = colon_death)

# 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)

# 2nd = extent
# aic_tbl2
death1.2 <- coxph(Surv(time, status) ~ rx+ node4 + extent, data = colon_death)

# 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)

# 3rd  = surg
# aic_tbl3
death1.3 <- coxph(Surv(time, status) ~ rx+ node4 + extent +
                      surg, data = colon_death)

# 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)),
        data = colon_death)
) |> set_names(uni_vars4)

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

```
# 4th = differ
# aic_tbl4
death1.4 <- coxph(Surv(time, status) ~ rx+ node4 + extent +
                      surg + differ, data = colon_death)

# 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
death1.5 <- coxph(Surv(time, status) ~ rx+ node4 + extent +
                      surg + differ + obstruct, data = colon_death)

deathAIC <- AIC(death1.1, death1.2, death1.3, death1.4, death1.5)
rownames(deathAIC) <- c("rx+node4", "rx+node4+extent", "rx+node4+extent+surg",
                        "rx+node4+extent+surg+differ", "rx+node4+extent+surg+differ+obstruct")
knitr::kable(deathAIC, caption = "Step AIC for Death")
```

Table 6: Step AIC for Death

	df	AIC
rx+node4	3	5446.809
rx+node4+extent	6	5432.636
rx+node4+extent+surg	7	5429.860
rx+node4+extent+surg+differ	9	5427.505
rx+node4+extent+surg+differ+obstruct	10	5425.768

After selecting `node4`, `node` was remove from the list of candidate covariate because they were highly correlated. 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 Coxph Model for Death

```
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 ($p < 2 \times 10^{-16}$), 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 $0.54 - 0.874$). `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 $1.02 - 1.55$).

Counting Process Model

The results from the marginal Cox proportional hazards models revealed that the combination treatment of Levamisole and 5-Fluorouracil (Lev+5FU) was significantly associated with reduced hazard for both recurrence and death. Additionally, time from initial surgery(surg), level of tumor spread(extent), and whether or not patients had more than four positive lymph nodes(node4) were important covariates associated with recurrence and death.

The data is looking at the time between recurrences of colon cancer and death. As a result, it cannot be treated as independent intervals like the gap model because there is a relationship between the recurrence of colon cancer and death from colon cancer. Thus, the counting process model is the best model to use.

Now the colon dataset counts the time between the beginning to the first recurrence, the time between the recurrence to the next recurrence or death.

However, the goal of the counting process model is to examine the effect of the treatments before recurrence and after recurrence. To do this, a new episode covariate must be created. If a given subject experienced a recurrence of colon cancer and died during the study, then they would have both an episode of 0 & episode of 1 for the respective rows. The same would apply if their recurrence occured and death was censored/outside of the study. However, if a subject did not have a recurrence of colon cancer but passed away in the study, then they would only have two episodes of 1, but the first row would be deleted. In the possibility that a patient didn't have a recurrence of colon cancer and their death was censored, there would be two episodes of 0, but the first row would remain.

Now that the counting process model is fully setup, it can be evaluated to see if it violates the cox proportional hazards assumption. The covariates from the marginal model will be included. Because of the episode covariate, the interaction between its levels and the treatment covariate will be tested.

```r
summary(coxph(Surv(start,stop,status) ~ rx*strata(episode) + strata(node4) +
                extent + surg, data = colon_counting))
```

```
## Call:
## coxph(formula = Surv(start, stop, status) ~ rx * strata(episode) + 
##     strata(node4) + extent + surg, data = colon_counting)
## 
##   n= 1328, number of events= 870 
## 
##                                        coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev                              -0.07550   0.92728  0.11019 -0.685   0.4933
## rxLev+5FU                          -0.51526   0.59734  0.12156 -4.239 2.25e-05
## extent2                             0.43753   1.54888  0.39801  1.099   0.2716
## extent3                             0.87331   2.39482  0.38163  2.288   0.0221
## extent4                             1.00335   2.72742  0.40739  2.463   0.0138
## surg1                               0.14892   1.16058  0.07503  1.985   0.0472
## rxLev:strata(episode)episode=1      0.21884   1.24463  0.15859  1.380   0.1676
## rxLev+5FU:strata(episode)episode=1  0.73645   2.08852  0.17310  4.255 2.09e-05
##                                       
## rxLev                                 
## rxLev+5FU                          ***
## extent2                               
## extent3                            *  
## extent4                            *  
## surg1                              *  
## rxLev:strata(episode)episode=1        
## rxLev+5FU:strata(episode)episode=1 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##                                  exp(coef) exp(-coef) lower .95 upper .95
## rxLev                               0.9273     1.0784    0.7472     1.151
## rxLev+5FU                           0.5973     1.6741    0.4707     0.758
## extent2                             1.5489     0.6456    0.7100     3.379
## extent3                             2.3948     0.4176    1.1335     5.060
## extent4                             2.7274     0.3666    1.2274     6.061
## surg1                               1.1606     0.8616    1.0019     1.344
## rxLev:strata(episode)episode=1      1.2446     0.8035    0.9121     1.698
## rxLev+5FU:strata(episode)episode=1    2.0885     0.4788    1.4876     2.932
##
## Concordance= 0.591  (se = 0.012 )
## Likelihood ratio test= 48.35  on 8 df,   p=8e-08
## Wald test            = 43.97  on 8 df,   p=6e-07
## Score (logrank) test = 44.86  on 8 df,   p=4e-07
```

From the summary, the model as a whole with the additional interaction between the given episode of colon cancer (time after recurrence vs time before recurrence) and the treatment covariate is statistically significant at a p-value of $8e-08$ vs $\alpha = 0.05$ for the critical value.

- Treatment: During the first episode *Lev+5FU* reduces the recurrence/death hazard approximately 40% with $HR = 0.5973$. Since the 95% confidence interval (0.4707, 0.758) does not include 1, the combination therapy can significantly reduce the hazard rate and increase the survival probability of patient.
  However, the significant interaction term *rxLev+5FU:strata(episode)episode=1* (95% CI: $1.4876-2.932$) shows this advantage disappears in the second episode which means after the recurrence (combined $HR = 0.5973 * 2.0885 \approx 1.25$). And *Lev* alone has hazard ratio 0.9273 reducing the hazard rate by 7.27% compared to the ‘obs‘. But such an effect is not significant in either episode. That is because the 95% confidence intervals before and after recurrence are (0.7472, 1.151) and (0.9121, 1.698), which include 1. Combining the interaction term
- *extent*: As the spread of the tumor developed from muscles to continguous structures, the hazard ratio to those who only had submucosa development increased to as high as 2.7274 times as likely to suffer a recurrence of colon cancer.
- *surg*: Patients with a long time from their initial surgery to registration in the study had a 16.06% greater hazard rate than those with a shorter time interval.

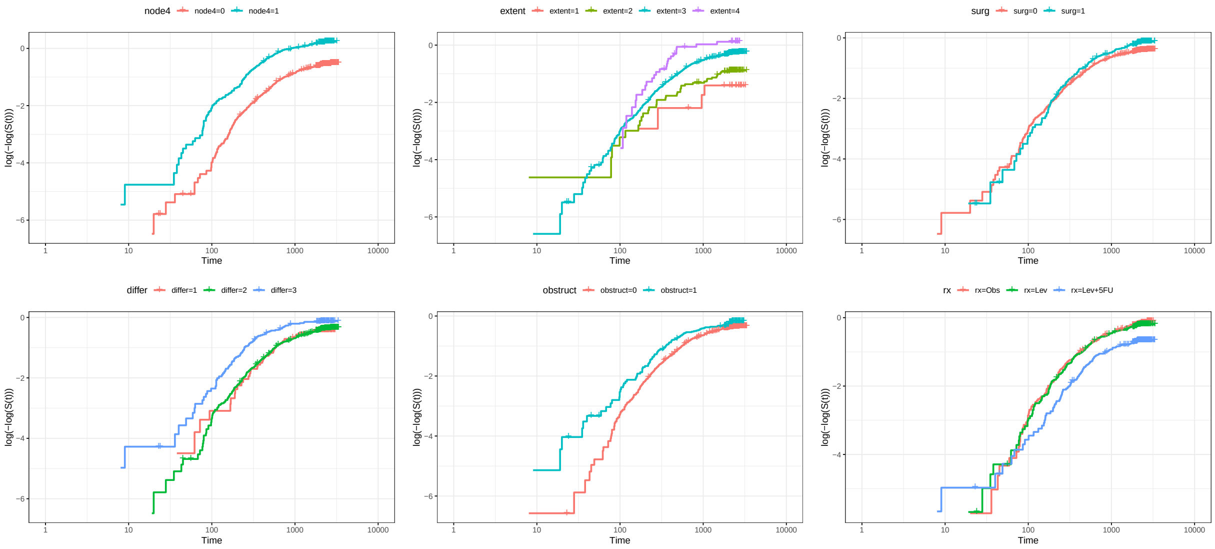
Proptional Hazards Assumption

To verify that the Cox models we use are valid, we assessed the proportional-hazards (PH) assumption. Our checking strategy combined (i) graphical inspection and (ii) formal statistical testing.

Recurrence Event

Log-log survival plots

Under the proportional-hazard (PH) assumption the category-specific curves should be roughly straight and parallel; systematic divergence, covergence, or corssing suggests non-PH behavior.



- *node4*: Two curves seems parallel and have no cross over time so *node4* satisfy the assumption.
- *extent*: Four curves cross each other at the early stage and remain parallel after that. Thus, *extent* satisfy the assumption.
- *surg*: Two curve cross each other at Time 250 so it might violate the assumption.
- *differ*: The green curve cross the red curve several time and the gap between blue curve and red curve decrease overtime. Thus, *differ* does not satisfy the assumption
- *obstruct* The gap between two curve seems to decrease over time. Thus, it violate the assumption.
- *rx*: Three curve cross each other at the early stage. Since there are few observations at the beginning, we ignore the crosses. And the red curve and the green curve are close so we ignore their crosses. And the gap between curve remain the same over time. Therefore, *rx* satisfy the assumption.

Cox ZPH

```
zph_recur <- cox.zph(full_recurrence)
print(zph_recur)
```

```
##         chisq df      p
## rx       0.40  2 0.8188
## node4   10.55  1 0.0012
## extent   1.11  3 0.7754
## surg     1.81  1 0.1786
## GLOBAL  13.73  7 0.0562
```

- *node4*: Since p-value is 0.0012 less than $\alpha = 0.05$, *node4* violate the assumption. This conclusion is different from the conclusion made from log-log plot.
- *extent*: Since p-value is 0.7754 greater than $\alpha = 0.05$, it satisfies the assumption.
- *surg*: p-value 0.1786 is large so it does not violate the assumption.
- *rx*: p-value 0.8188 is large so it does not violate the assumption.

Revised Coxph Recurrence Model

`node4` obtains different conclusion from log-log plot and Cox ZPH test regarding the violation of PH assumption. Since Cox ZPH is a more powerful statistical test, we conclude that `node4` violate the PH assumption. Thus, the coxph model for recurrence event has one covariates `node4` violate the assumption.

```
revised_recurrence <- coxph(Surv(time, status) ~ strata(node4) + extent + surg
                            + rx, data = recurrence_data)
cox.zph(revised_recurrence)
```

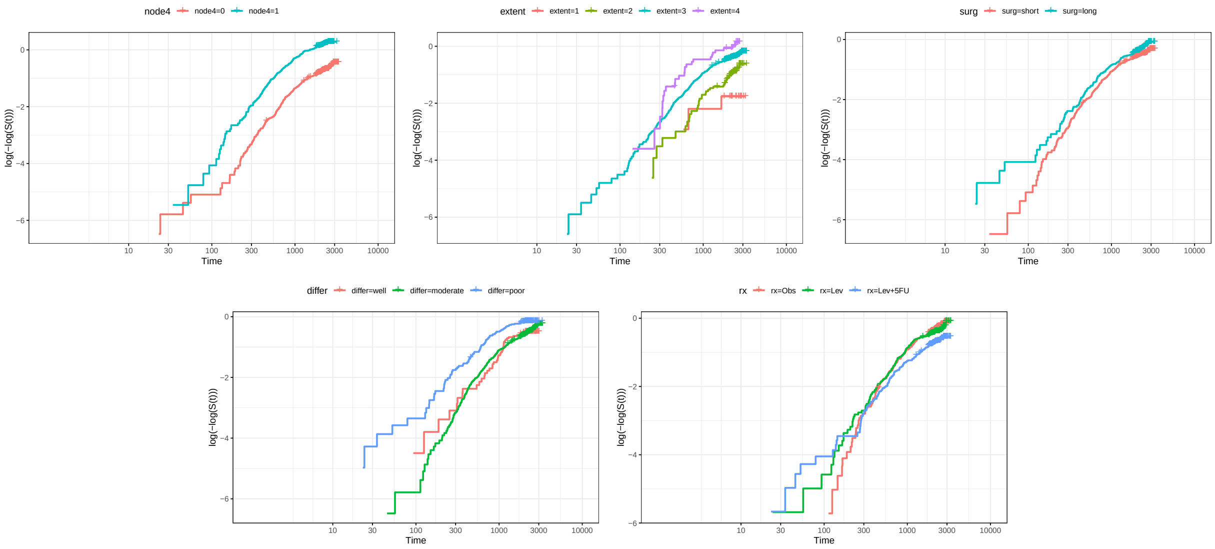
```
##        chisq df    p
## extent 1.340  3 0.72
## surg   1.518  1 0.22
## rx     0.364  2 0.83
## GLOBAL 3.120  6 0.79
```

After stratifying `node4`, the output of `cox.zph` for revised cox proportional hazard model of recurrence shows that all covariates, `extent`, `surg`, and `rx`, satisfy the PH assumption.

Death Event

Log-log survival plots

Cox proportional hazard model for death event use log-log plot to check the PH assumption based on the same criteria in the recurrence event.



- *node4*: Two curves seems parallel so *node4* satisfy the assumption.
- *extent*: At the early stage the blue curve cross the purple curve and the green curve cross the red curve. Because the purple curve and the red curve have few observations, we ignore the effect of crosses. After that, the purple, blue, and green curves are relatively parallel but the red curve seems to deviate from them. Thus, *extent* satisfy assumption.
- *surg*: Two curve never cross and remain parallel after Time 200. So, *surg* satisfy the assumption.

- *differ*: The green curve cross the other two curves four times over time. And the gap between each curve become smaller over time. Thus, *differ* violate the assumption.
- *rx*: Three curve cross each other at the early stage but with few observations. And the green and red curve are closed to each other but the gap between each curve is relatively parallel. It is hard to make conclusion by looking at the log-log plot so we will use the statistical testing to test assumption.

Cox ZPH

```
zph_death <- cox.zph(full_death)
print(zph_death)
```

```
##          chisq df       p
## node4   5.4346  1 0.01974
## extent  6.5521  3 0.08763
## surg    0.0205  1 0.88616
## differ 15.8513  2 0.00036
## rx      2.5680  2 0.27693
## GLOBAL 27.5832  9 0.00112
```

- *node4*: Since p-value is 0.01974 less than $\alpha = 0.05$, *node4* violate the assumption. This conclusion is different from the conclusion made from log-log plot.
- *extent*: Since p-value is 0.08763 greater than $\alpha = 0.05$, it does not violate the assumption. Although there is no sufficient evidence of non-PH, it is concerning about the assumption.
- *surg*: p-value 0.88616 is large so it does not violate the assumption.
- *differ*: Since p-value is 0.00036 smaller than $\alpha = 0.05$, it violate the assumption.
- *rx*: p-value 0.27693 is large so it does not violate the assumption.

Revised Coxph Death Model

The cox proportional hazard model for death event has two covariates (`node4`, `differ`) violating the PH assumption. In fact, `node4` are correlated because poorly differentiated tumors exhibit less resemblance and may grow more rapidly which tend to have more nodes. Also, `node4` has two levels (0 and 1) less than `differ` which have three levels (1, 2, and 3). Since stratifying a covariate with more level could over complicate the model, the model will exclude `differ` and stratify `node4`.

```
revised_death <- coxph(Surv(time, status) ~ strata(node4) + extent +
                      surg + rx, data = colon_death)
# summary(revised_death)
cox.zph(revised_death)
```

```
##         chisq df    p
## extent 5.9454  3 0.11
## surg   0.0432  1 0.84
## rx     2.2081  2 0.33
## GLOBAL 8.5036  6 0.20
```

After deleting `differ` and stratifying `node4`, the output of `cox.zph` for revised cox proportional hazard model of recurrence shows that all covariates, `extent`, `surg`, and `rx`, satisfy the PH assumption.

Conclusion

The results from the marginal and counting process models found that the combination of both Levamisole and Fluoracil has a statistically significant effect in reducing the hazard rate of a recurrence or death from colon cancer. In the marginal model for the deaths from colon cancer, the hazard ratio for the combined treatment was 0.689, so patients who took this combined treatment of Levamisole and Fluoracil were 31% less likely to die from colon cancer compared to patients who took no treatment at all. The 95% confidence interval for the true hazard ratio between the combined treatment and no treatment levels was (0.54, 0.874). In the marginal model for the recurrences from colon cancer, the hazard ratio for the combined treatment was 0.6107, so patients who took this combined treatment of Levamisole and Fluoracil were $\approx$ 39% less likely to experience a recurrence of colon cancer compared to patients who took no treatment at all. The 95% confidence interval for the true hazard ratio between the combined treatment and no treatment levels was (0.4811, 0.7752). In the counting process model for all of the observations of colon cancer, recurrences plus deaths, the hazard ratio for the combined treatment was 0.5973, so patients who took this combined treatment of Levamisole and Fluoracil were $\approx$ 40% less likely to experience a recurrence or death from colon cancer compared to patients who took no treatment at all. The 95% confidence interval for the true hazard ratio between the combined treatment and no treatment levels was (0.4707, 0.758).

References

1. colon: Chemotherapy for Stage B/C colon cancer. Colon Cancer. Retrieved May 1st, 2025, from https://rdrr.io/cran/survival/man/colon.html

2. CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, CM Tangen, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH Veeder and JA Maillard, Fluorouracil plus Levamisole as an effective adjuvant therapy after resection of stage II colon carcinoma: a final report. Annals of Internal Med, 122:321-326, 1991.