P8108 Final Project

Group 6

2025-11-16

Data Import

```
lung_df <- survival::lung %>%
  janitor::clean_names() %>%
  mutate(
    inst = as.factor(inst),
    time = as.numeric(time),
    status = as.factor(status),
    event = status == 2,
    age = as.numeric(age),
    sex = factor(sex, levels = c(1, 2), labels = c("Male", "Female")),
    ph_ecog = factor(ph_ecog, ordered = TRUE),
    ph_karno = as.numeric(ph_karno),
    pat_karno = as.numeric(pat_karno),
    meal_cal = as.numeric(meal_cal),
    wt_loss = as.numeric(wt_loss)
  )
```

We create a new variable called `event` to indicate survival status, where 1 represents death and 0 represents censoring.

The variable `ph_ecog` (ECOG performance score, 0–3) is treated as an ordinal variable. For descriptive and Kaplan–Meier analyses, it is handled as a categorical factor to visualize group differences. It might be modeled as an ordinal numeric variable in Cox model.

Check NAs

```
summary(lung_df)
```

```
##       inst          time        status       age            sex      ph_ecog
##  1      : 36   Min.   :   5.0   1: 63   Min.   :39.00   Male  :138   0   : 63
##  12     : 23   1st Qu.: 166.8   2:165   1st Qu.:56.00   Female: 90   1   :113
##  13     : 20   Median : 255.5           Median :63.00                2   : 50
##  3      : 19   Mean   : 305.2           Mean   :62.45                3   :  1
##  11     : 18   3rd Qu.: 396.5           3rd Qu.:69.00                NA's:  1
##  (Other):111   Max.   :1022.0           Max.   :82.00
##  NA's   :  1
##     ph_karno        pat_karno        meal_cal        wt_loss
##  Min.   : 50.00   Min.   : 30.00   Min.   :  96.0   Min.   :-24.000
##  1st Qu.: 75.00   1st Qu.: 70.00   1st Qu.: 635.0   1st Qu.:  0.000
##  Median : 80.00   Median : 80.00   Median : 975.0   Median :  7.000
##  Mean   : 81.94   Mean   : 79.96   Mean   : 928.8   Mean   :  9.832
##  3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1150.0   3rd Qu.: 15.750
```

```
##  Max.   :100.00   Max.   :100.00   Max.   :2600.0   Max.   : 68.000
##  NA's   :1        NA's   :3        NA's   :47       NA's   :14
##    event
##  Mode :logical
##  FALSE:63
##  TRUE :165
##
##
##
##
```

```
lung_cc <- lung_df %>%
  filter(complete.cases(time, event, age, sex, ph_ecog,
                        ph_karno, pat_karno, meal_cal, wt_loss, inst))
summary(lung_cc)
```

```
##       inst          time         status       age            sex      ph_ecog
##  1      :28   Min.   :   5.0   1: 47   Min.   :39.00   Male  :103   0:47
##  12     :16   1st Qu.: 174.5   2:120   1st Qu.:57.00   Female: 64   1:81
##  11     :13   Median : 268.0           Median :64.00                2:38
##  13     :13   Mean   : 309.9           Mean   :62.57                3: 1
##  22     :13   3rd Qu.: 419.5           3rd Qu.:70.00
##  3      :12   Max.   :1022.0           Max.   :82.00
##  (Other):72
##     ph_karno        pat_karno         meal_cal         wt_loss
##  Min.   : 50.00   Min.   : 30.00   Min.   :  96.0   Min.   :-24.000
##  1st Qu.: 70.00   1st Qu.: 70.00   1st Qu.: 619.0   1st Qu.:  0.000
##  Median : 80.00   Median : 80.00   Median : 975.0   Median :  7.000
##  Mean   : 82.04   Mean   : 79.58   Mean   : 929.1   Mean   :  9.719
##  3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1162.5   3rd Qu.: 15.000
##  Max.   :100.00   Max.   :100.00   Max.   :2600.0   Max.   : 68.000
##
##    event
##  Mode :logical
##  FALSE:47
##  TRUE :120
##
##
##
##
```

There are NAs in this data, so we will also create another dataset that observations with missing values will be excluded to ensure that all variables used in the analysis had complete information.

Both the original dataset and the complete-case dataset were retained for further analyses to allow comparisons and sensitivity checks.

EDA

Descriptive Table

```
lung_cc %>%
  select(time, event, age, sex, ph_ecog, ph_karno, pat_karno, meal_cal, wt_loss) %>%
  tbl_summary(
    by = sex,
    missing = "no",
```

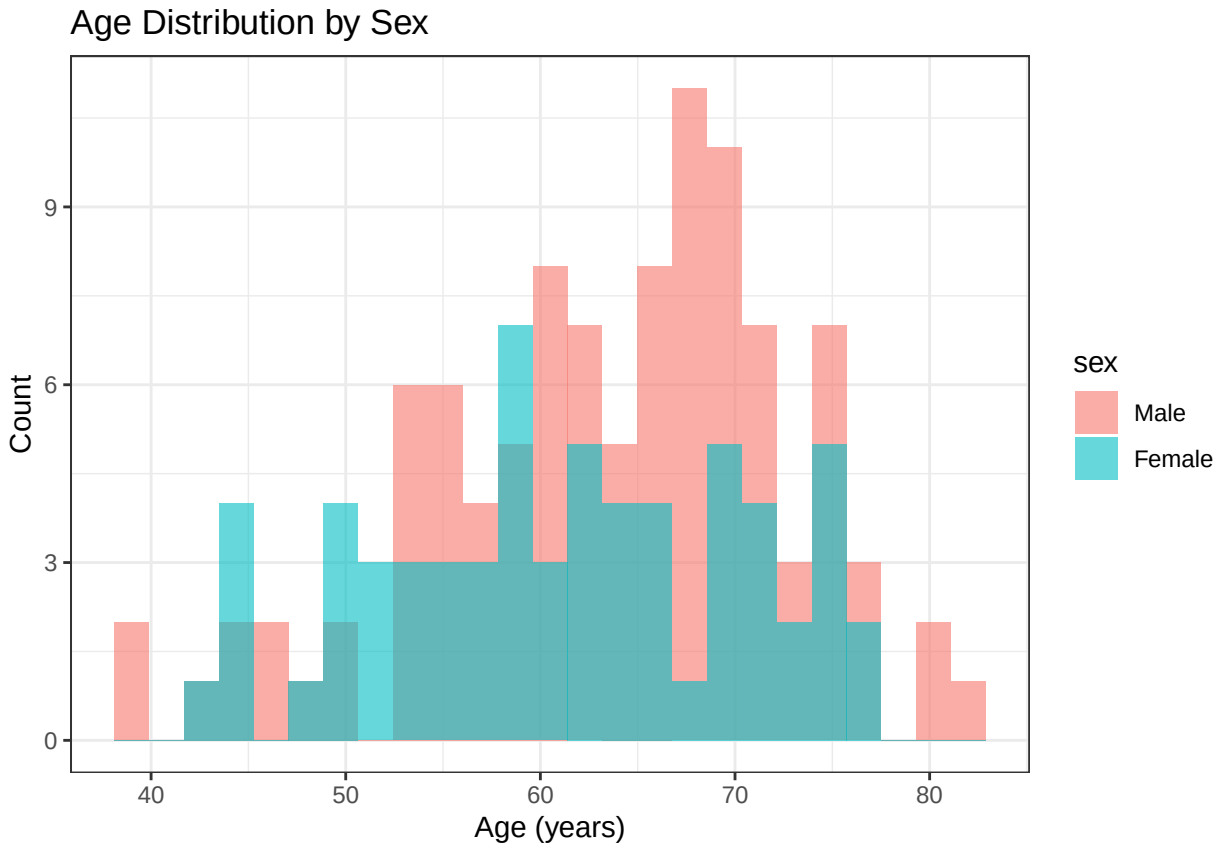
Characteristic	Overall N = 167[1]	Male N = 103[1]	Female N = 64[1]	p-value[2]
Survival time (days)	310 (209)	291 (208)	340 (209)	0.069
Death indicator	120 (72%)	82 (80%)	38 (59%)	0.005
Age (years)	63 (9)	63 (9)	61 (9)	0.10
ECOG performance status				>0.9
0	47 (28%)	28 (27%)	19 (30%)	
1	81 (49%)	52 (50%)	29 (45%)	
2	38 (23%)	22 (21%)	16 (25%)	
3	1 (0.6%)	1 (1.0%)	0 (0%)	
Physician Karnofsky score				0.8
50	4 (2.4%)	3 (2.9%)	1 (1.6%)	
60	16 (9.6%)	8 (7.8%)	8 (13%)	
70	24 (14%)	16 (16%)	8 (13%)	
80	47 (28%)	28 (27%)	19 (30%)	
90	50 (30%)	32 (31%)	18 (28%)	
100	26 (16%)	16 (16%)	10 (16%)	
Patient Karnofsky score				0.3
30	2 (1.2%)	1 (1.0%)	1 (1.6%)	
40	2 (1.2%)	1 (1.0%)	1 (1.6%)	
50	3 (1.8%)	2 (1.9%)	1 (1.6%)	
60	23 (14%)	15 (15%)	8 (13%)	
70	30 (18%)	24 (23%)	6 (9.4%)	
80	37 (22%)	20 (19%)	17 (27%)	
90	44 (26%)	24 (23%)	20 (31%)	
100	26 (16%)	16 (16%)	10 (16%)	
Meal calories	929 (413)	985 (428)	840 (374)	0.027
Weight loss (kg)	10 (13)	12 (13)	7 (13)	0.015

[1]Mean (SD); n (%)
[2]Wilcoxon rank sum test

```
    statistic = list(all_continuous() ~ "{mean} ({sd})", all_categorical() ~ "{n} ({p}%)"),
    label = list(
      time ~ "Survival time (days)",
      event ~ "Death indicator",
      age ~ "Age (years)",
      ph_ecog ~ "ECOG performance status",
      ph_karno ~ "Physician Karnofsky score",
      pat_karno ~ "Patient Karnofsky score",
      meal_cal ~ "Meal calories",
      wt_loss ~ "Weight loss (kg)"
    )
  ) %>%
  add_overall() %>%
  add_p(test = everything() ~ "wilcox.test") %>%
  bold_labels()
```

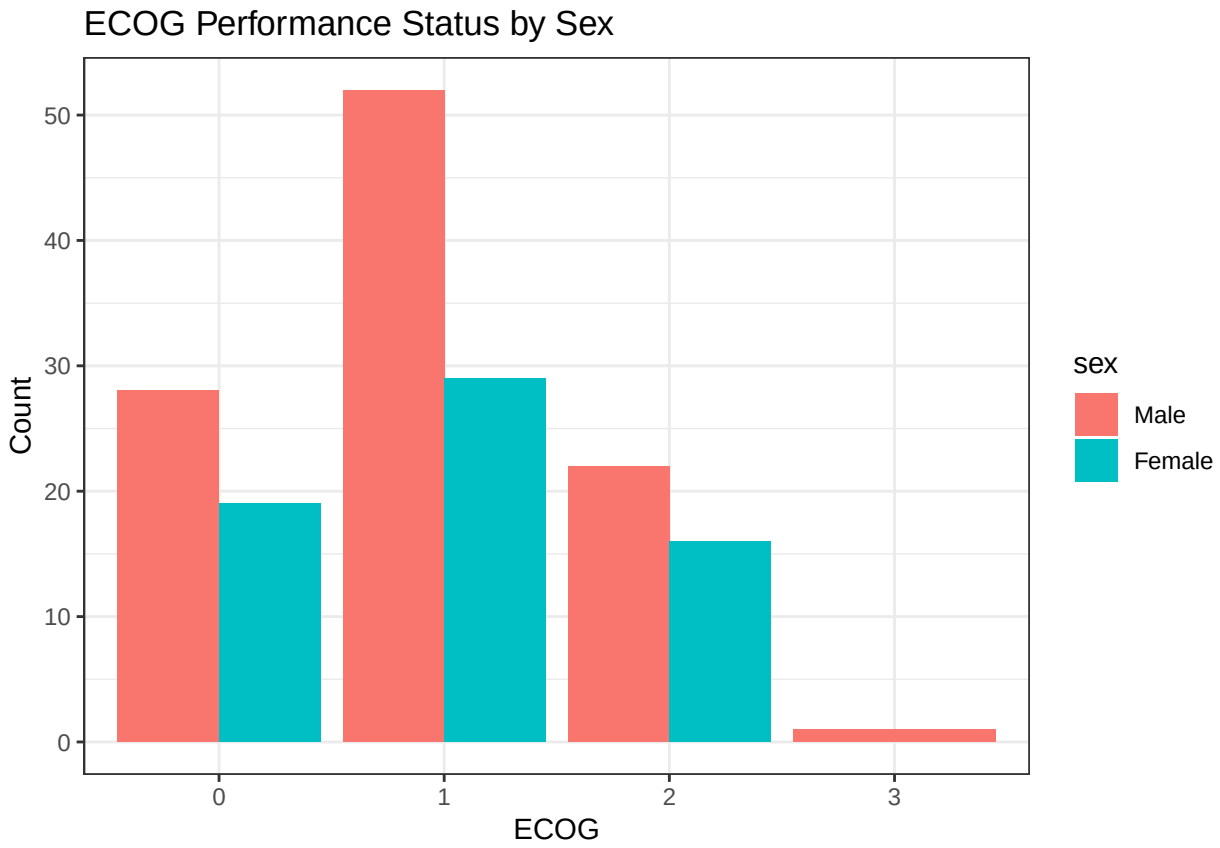
Age Distribution by Sex

```
ggplot(lung_cc, aes(x = age, fill = sex)) +
  geom_histogram(bins = 25, position = "identity", alpha = 0.6) +
  theme_bw() +
  labs(title = "Age Distribution by Sex", x = "Age (years)", y = "Count")
```



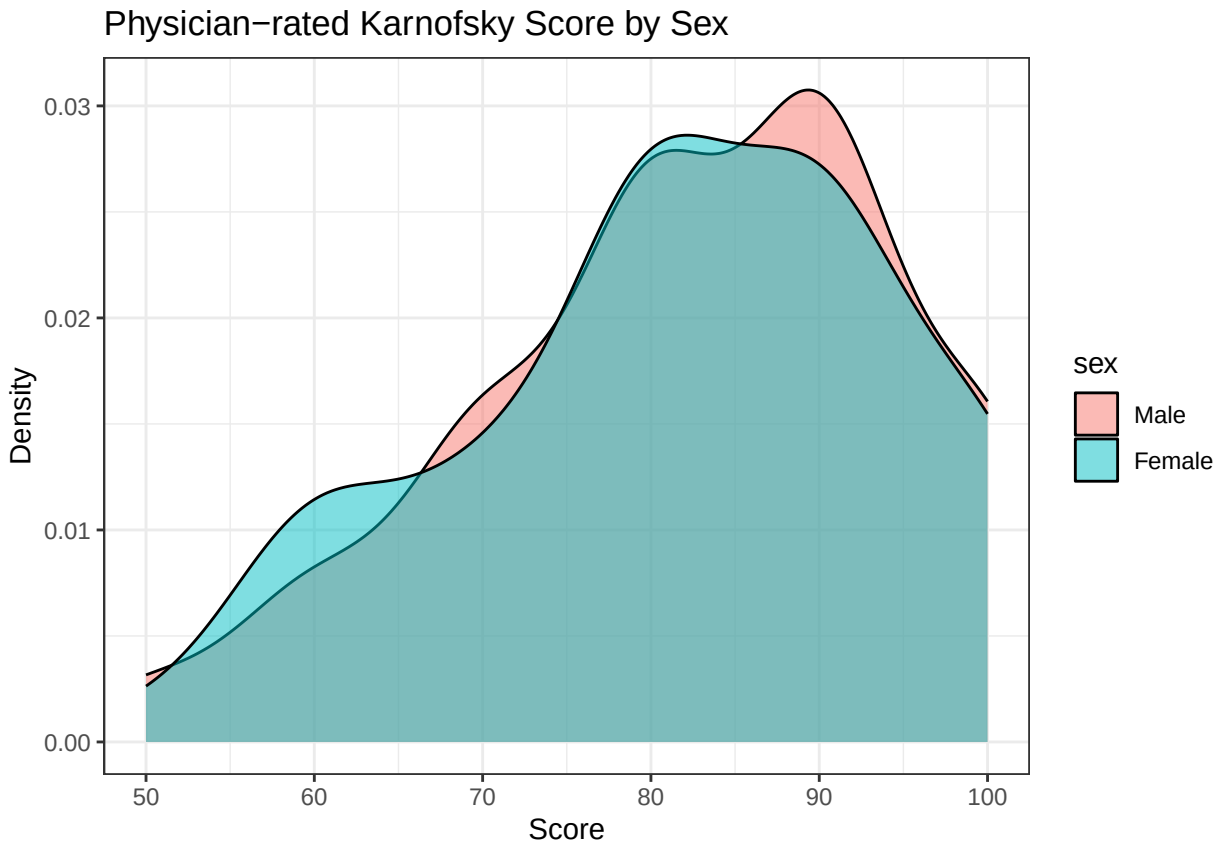
ECOG Performance Status by Sex

```
ggplot(lung_cc, aes(x = ph_ecog, fill = sex)) +
  geom_bar(position = "dodge") +
  theme_bw() +
  labs(title = "ECOG Performance Status by Sex", x = "ECOG", y = "Count")
```



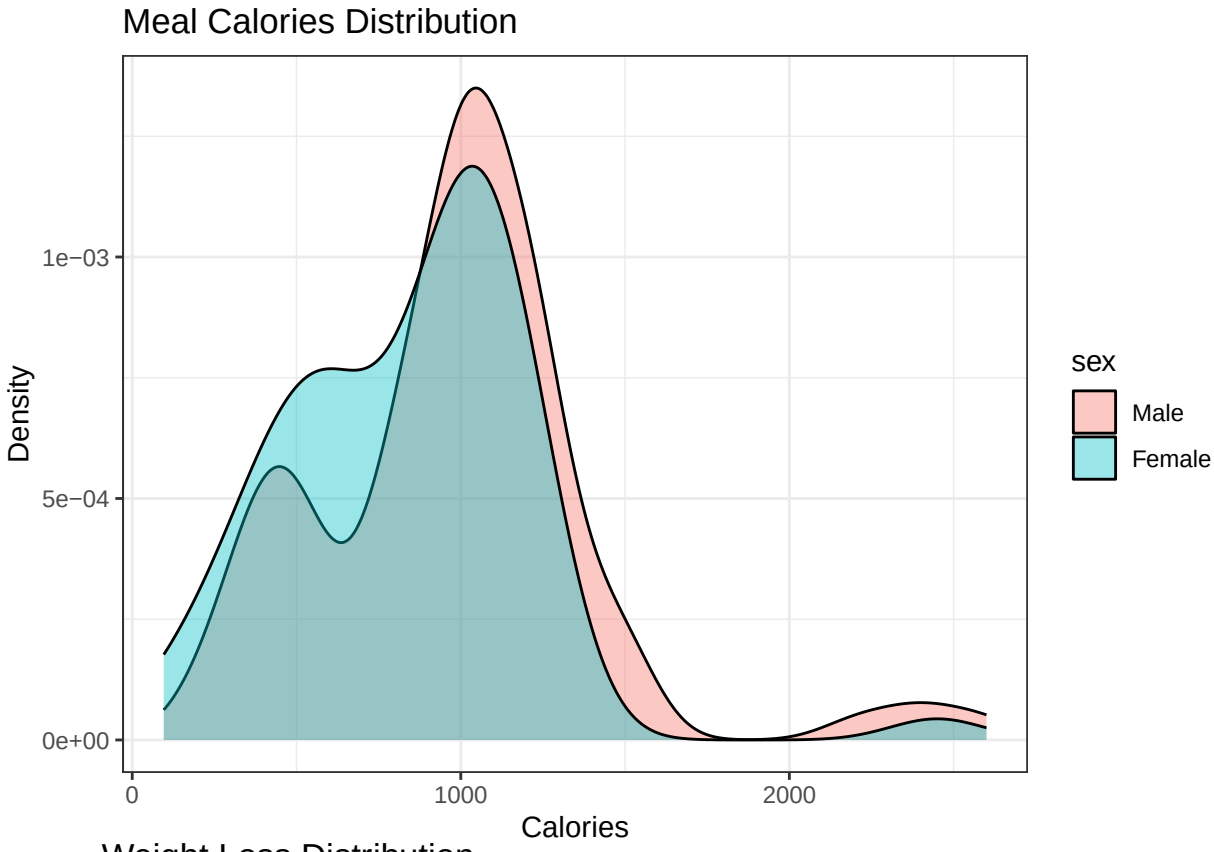
Physician-rated Karnofsky Score by Sex

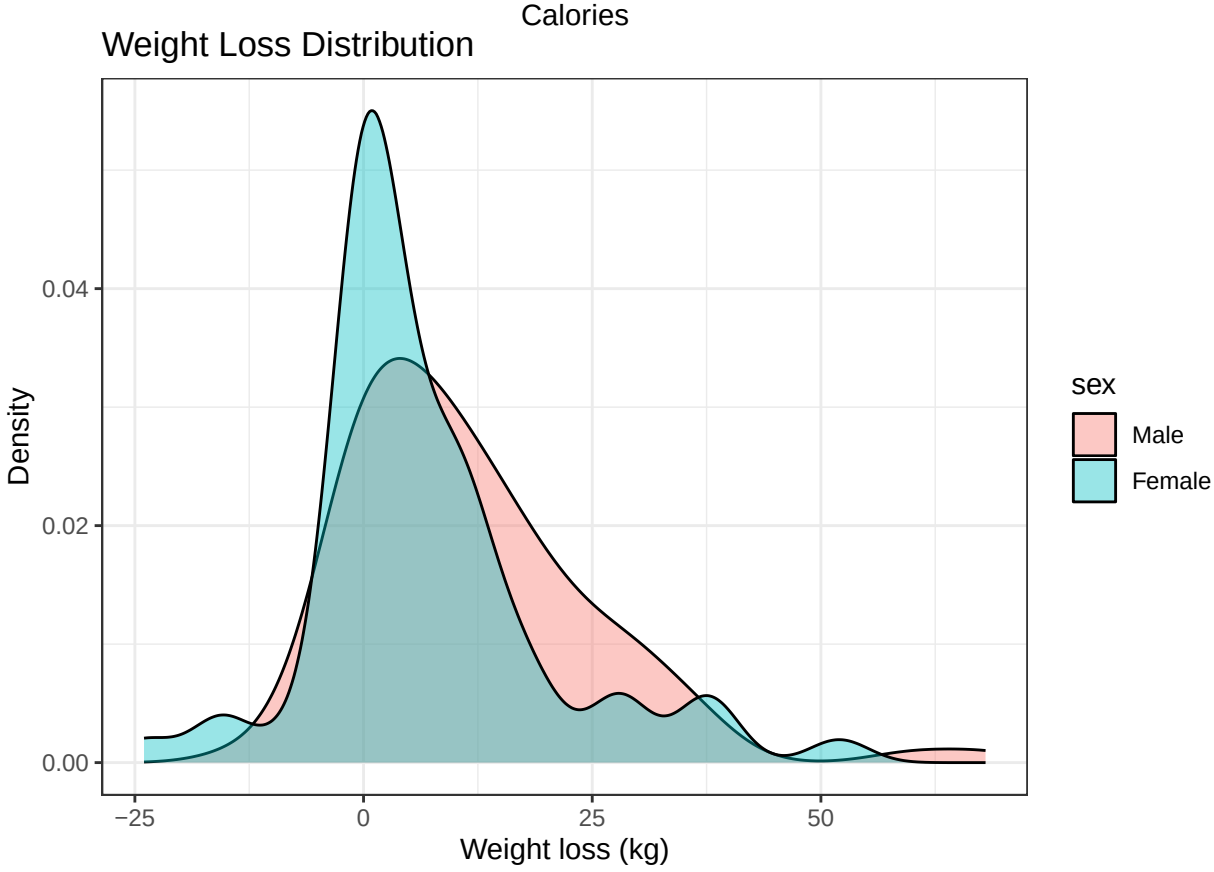
```
ggplot(lung_cc, aes(x = ph_karno, fill = sex)) +
  geom_density(alpha = 0.5) +
  theme_bw() +
  labs(title = "Physician-rated Karnofsky Score by Sex", x = "Score", y = "Density")
```



Meal Calories and Weight Loss Distributions

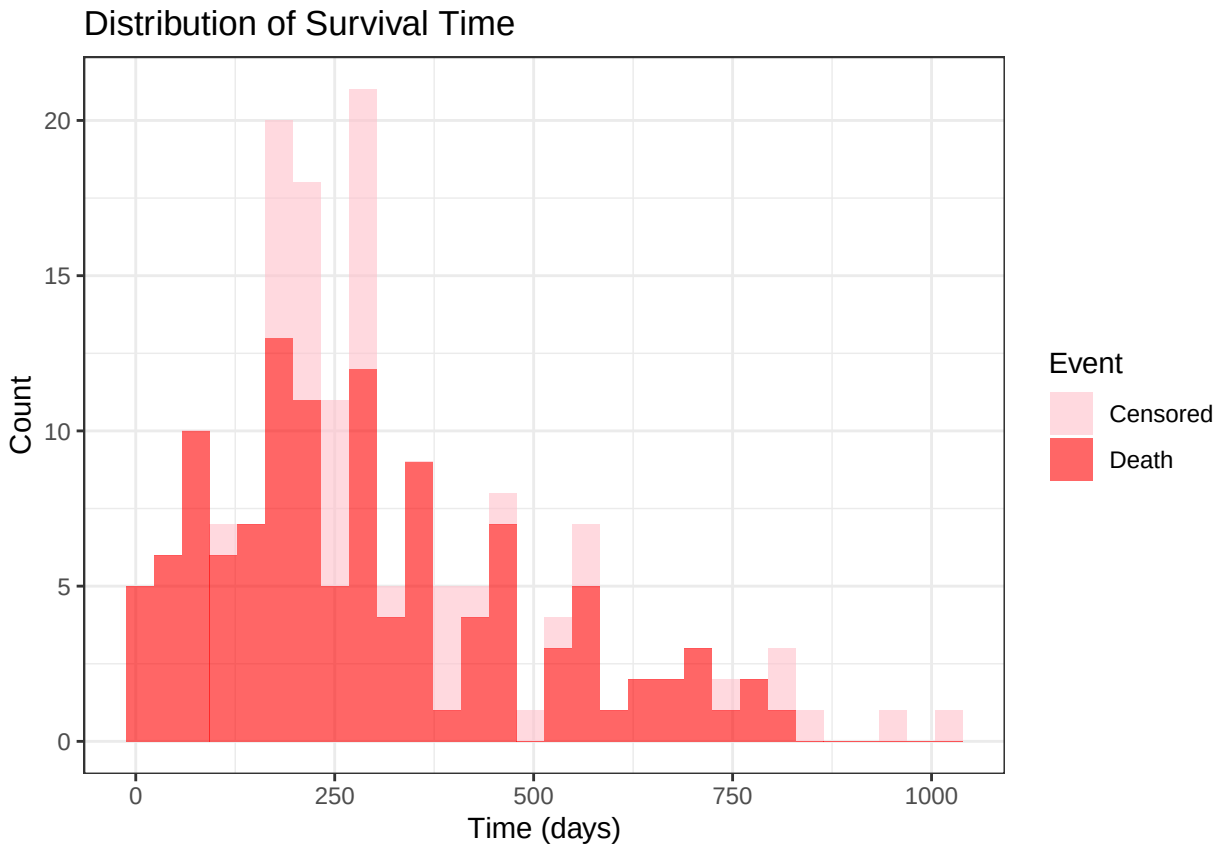
```
p1 <- ggplot(lung_cc, aes(x = meal_cal, fill = sex)) +
  geom_density(alpha = 0.4) +
  labs(title = "Meal Calories Distribution", x = "Calories", y = "Density") +
  theme_bw()

p2 <- ggplot(lung_cc, aes(x = wt_loss, fill = sex)) +
  geom_density(alpha = 0.4) +
  labs(title = "Weight Loss Distribution", x = "Weight loss (kg)", y = "Density") +
  theme_bw()

p1; p2
```


Meal Calories Distribution
Density
1e−03
5e−04
0e+00
0
1000
2000
Calories
sex
Male
Female


Weight Loss Distribution
Density
0.04
0.02
0.00
−25
0
25
50
Weight loss (kg)
sex
Male
Female

Distribution of Survival Time

```
ggplot(lung_cc, aes(x = time, fill = event)) +
  geom_histogram(bins = 30, alpha = 0.6) +
  theme_bw() +
  labs(title = "Distribution of Survival Time", x = "Time (days)", y = "Count") +
  scale_fill_manual(values = c("PINK", "RED"), name = "Event", labels = c("Censored", "Death"))
```



Comparison of Survival Curves

```
# Create a survival object, which bundles the time and event data
surv_object <- Surv(time = lung_cc$time, event = lung_cc$event)
```

```
# Fit the Overall Kaplan-Meier Model
km_overall_fit <- survfit(surv_object ~ 1, data = lung_cc)

print(km_overall_fit)
```

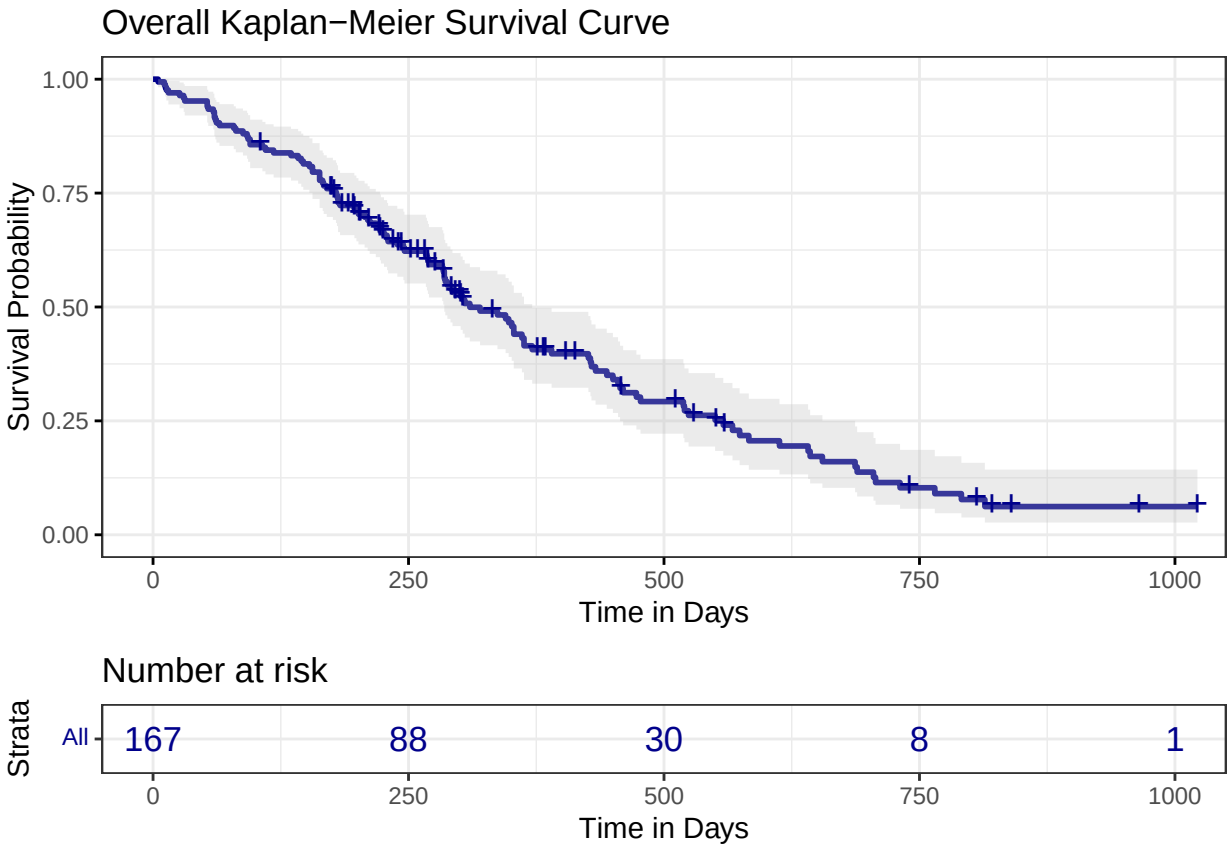
```
## Call: survfit(formula = surv_object ~ 1, data = lung_cc)
##
##         n events median 0.95LCL 0.95UCL
## [1,] 167    120    310     285     371
```

```
ggsurvplot(
  km_overall_fit,
  data = lung_cc,
  conf.int = TRUE,
```

```
  risk.table = TRUE,
  risk.table.col = "strata",
  title = "Overall Kaplan-Meier Survival Curve",
  xlab = "Time in Days",
  ylab = "Survival Probability",
  ggtheme = theme_bw(),
  palette = "darkblue",
  legend = "none"
)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
```



Sex

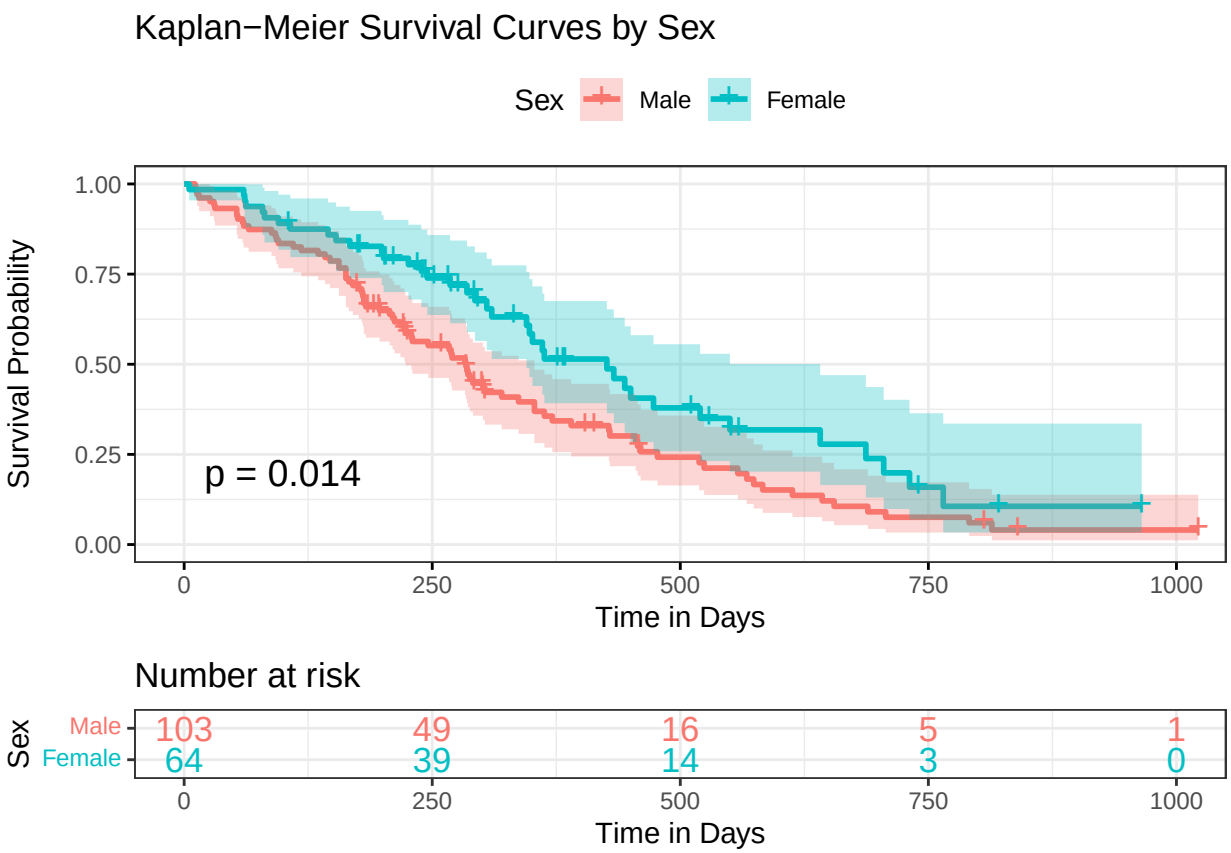
```
# Fit the Kaplan-Meier model to estimate survival curves for each group ('sex')
km_sex_fit <- survfit(surv_object ~ sex, data = lung_cc)

summary(km_sex_fit)
```

```
## Call: survfit(formula = surv_object ~ sex, data = lung_cc)
##
##                 sex=Male
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    11    103       1   0.9903 0.00966       0.9715        1.000
##    12    102       1   0.9806 0.01360       0.9543        1.000
##    13    101       1   0.9709 0.01657       0.9389        1.000
##    15    100       1   0.9612 0.01904       0.9246        0.999
##    26     99       1   0.9515 0.02118       0.9108        0.994
##    30     98       1   0.9417 0.02308       0.8976        0.988
##    31     97       1   0.9320 0.02480       0.8847        0.982
##    53     96       2   0.9126 0.02782       0.8597        0.969
##    54     94       1   0.9029 0.02917       0.8475        0.962
##    59     93       1   0.8932 0.03043       0.8355        0.955
##    60     92       1   0.8835 0.03161       0.8237        0.948
##    65     91       1   0.8738 0.03272       0.8119        0.940
##    88     90       1   0.8641 0.03377       0.8004        0.933
##    92     89       1   0.8544 0.03476       0.7889        0.925
##    93     88       1   0.8447 0.03569       0.7775        0.918
##    95     87       1   0.8350 0.03658       0.7663        0.910
##   110     86       1   0.8252 0.03742       0.7551        0.902
##   118     85       1   0.8155 0.03822       0.7440        0.894
##   135     84       1   0.8058 0.03898       0.7329        0.886
##   142     83       1   0.7961 0.03970       0.7220        0.878
##   147     82       1   0.7864 0.04038       0.7111        0.870
##   156     81       2   0.7670 0.04165       0.6895        0.853
##   163     79       3   0.7379 0.04333       0.6576        0.828
##   166     76       1   0.7282 0.04384       0.6471        0.819
##   170     75       1   0.7184 0.04432       0.6366        0.811
##   176     73       1   0.7086 0.04479       0.6260        0.802
##   179     72       1   0.6988 0.04523       0.6155        0.793
##   180     71       1   0.6889 0.04566       0.6050        0.784
##   181     70       2   0.6692 0.04642       0.5842        0.767
##   183     68       1   0.6594 0.04677       0.5738        0.758
##   197     64       1   0.6491 0.04716       0.5629        0.748
##   207     62       1   0.6386 0.04755       0.5519        0.739
##   210     61       1   0.6282 0.04791       0.5409        0.729
##   212     60       1   0.6177 0.04824       0.5300        0.720
##   218     59       1   0.6072 0.04855       0.5191        0.710
##   222     57       1   0.5966 0.04885       0.5081        0.700
##   223     55       1   0.5857 0.04915       0.4969        0.690
##   229     52       1   0.5745 0.04948       0.4852        0.680
##   230     51       1   0.5632 0.04977       0.4736        0.670
##   246     50       1   0.5519 0.05004       0.4621        0.659
##   267     48       1   0.5404 0.05030       0.4503        0.649
##   269     47       1   0.5289 0.05053       0.4386        0.638
##   270     46       1   0.5174 0.05072       0.4270        0.627
```

```
##   283     45       1   0.5059 0.05088       0.4154        0.616
##   284     44       1   0.4944 0.05101       0.4039        0.605
##   285     42       1   0.4827 0.05113       0.3922        0.594
##   286     41       1   0.4709 0.05122       0.3805        0.583
##   288     40       1   0.4591 0.05128       0.3689        0.571
##   291     39       1   0.4473 0.05129       0.3573        0.560
##   301     36       1   0.4349 0.05135       0.3451        0.548
##   303     34       1   0.4221 0.05141       0.3325        0.536
##   320     32       1   0.4089 0.05147       0.3195        0.523
##   337     31       1   0.3957 0.05147       0.3067        0.511
##   353     30       2   0.3694 0.05131       0.2813        0.485
##   363     28       1   0.3562 0.05114       0.2688        0.472
##   371     27       1   0.3430 0.05092       0.2564        0.459
##   390     26       1   0.3298 0.05064       0.2441        0.446
##   428     23       1   0.3154 0.05043       0.2306        0.432
##   429     22       1   0.3011 0.05014       0.2173        0.417
##   455     21       1   0.2868 0.04976       0.2041        0.403
##   457     20       1   0.2724 0.04929       0.1911        0.388
##   460     18       1   0.2573 0.04882       0.1774        0.373
##   477     17       1   0.2422 0.04824       0.1639        0.358
##   519     16       1   0.2270 0.04754       0.1506        0.342
##   524     15       1   0.2119 0.04672       0.1375        0.326
##   558     14       1   0.1968 0.04577       0.1247        0.310
##   567     13       1   0.1816 0.04468       0.1121        0.294
##   574     12       1   0.1665 0.04344       0.0998        0.278
##   583     11       1   0.1514 0.04205       0.0878        0.261
##   613     10       1   0.1362 0.04048       0.0761        0.244
##   643      9       1   0.1211 0.03870       0.0647        0.227
##   655      8       1   0.1059 0.03671       0.0537        0.209
##   689      7       1   0.0908 0.03444       0.0432        0.191
##   707      6       1   0.0757 0.03185       0.0332        0.173
##   791      5       1   0.0605 0.02886       0.0238        0.154
##   814      3       1   0.0404 0.02533       0.0118        0.138
##
##                 sex=Female
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5     64       1    0.984  0.0155       0.9545        1.000
##    60     63       1    0.969  0.0217       0.9270        1.000
##    61     62       1    0.953  0.0264       0.9027        1.000
##    62     61       1    0.938  0.0303       0.8800        0.999
##    79     60       1    0.922  0.0335       0.8584        0.990
##    81     59       1    0.906  0.0364       0.8376        0.981
##    95     58       1    0.891  0.0390       0.8174        0.970
##   107     56       1    0.875  0.0414       0.7972        0.960
##   145     55       1    0.859  0.0436       0.7774        0.949
##   153     54       1    0.843  0.0456       0.7581        0.937
##   167     53       1    0.827  0.0475       0.7390        0.925
##   199     50       1    0.810  0.0493       0.7194        0.913
##   201     49       1    0.794  0.0510       0.7000        0.900
##   226     45       1    0.776  0.0528       0.6794        0.887
##   239     43       1    0.758  0.0546       0.6584        0.873
##   245     40       1    0.739  0.0564       0.6366        0.859
##   268     37       1    0.719  0.0583       0.6136        0.843
##   285     34       1    0.698  0.0603       0.5894        0.827
```

```
##   293     32       1    0.676  0.0623       0.5647        0.810
##   305     30       1    0.654  0.0641       0.5394        0.792
##   310     29       1    0.631  0.0658       0.5146        0.774
##   345     27       1    0.608  0.0674       0.4892        0.755
##   348     26       1    0.584  0.0687       0.4642        0.736
##   351     25       1    0.561  0.0698       0.4397        0.716
##   361     24       1    0.538  0.0707       0.4155        0.696
##   363     23       1    0.514  0.0714       0.3918        0.675
##   426     19       1    0.487  0.0726       0.3639        0.653
##   433     18       1    0.460  0.0734       0.3366        0.629
##   444     17       1    0.433  0.0739       0.3100        0.605
##   450     16       1    0.406  0.0741       0.2839        0.581
##   473     15       1    0.379  0.0739       0.2585        0.556
##   520     13       1    0.350  0.0738       0.2314        0.529
##   550     11       1    0.318  0.0736       0.2020        0.501
##   641      8       1    0.278  0.0744       0.1648        0.470
##   687      7       1    0.239  0.0736       0.1303        0.437
##   705      6       1    0.199  0.0713       0.0984        0.401
##   731      5       1    0.159  0.0672       0.0695        0.364
##   765      3       1    0.106  0.0623       0.0335        0.335
```

```r
# --- Visualize the Estimated Survival Probabilities ---

# Generate the Kaplan-Meier plot using ggsurvplot
ggsurvplot(
  km_sex_fit,
  data = lung_cc,
  pval = TRUE,                # The p-value from the log-rank test will be displayed
  conf.int = TRUE,            # Display confidence intervals for the curves
  risk.table = TRUE,          # Add a table showing the number of subjects at risk
  risk.table.col = "strata",  # Color the risk table to match the curves
  legend.labs = c("Male", "Female"),
  legend.title = "Sex",
  xlab = "Time in Days",
  ylab = "Survival Probability",
  title = "Kaplan-Meier Survival Curves by Sex",
  ggtheme = theme_bw()        # Apply a clean theme
)
```




```
# With rho = 0 this is the log-rank or Mantel-Haenszel test, and with rho = 1 it is equivalent to the P
log_rank_sex <- survdiff(surv_object ~ sex, data = lung_cc, rho = 0)
print(log_rank_sex)
```

```
## Call:
## survdiff(formula = surv_object ~ sex, data = lung_cc, rho = 0)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Male   103       82     68.7      2.57      6.05
## sex=Female  64       38     51.3      3.44      6.05
##
##  Chisq= 6  on 1 degrees of freedom, p= 0.01
```

```
wilcoxon_sex <- survdiff(surv_object ~ sex, data = lung_cc, rho = 1)
print(wilcoxon_sex)
```

```
## Call:
## survdiff(formula = surv_object ~ sex, data = lung_cc, rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Male   103     51.0     41.8      2.01      6.76
## sex=Female  64     21.1     30.3      2.77      6.76
##
##  Chisq= 6.8  on 1 degrees of freedom, p= 0.009
```

The p-value is even smaller than the log-rank p-value. This not only confirms the result of the log-rank test but also suggests that the survival advantage for females is particularly pronounced at the beginning and middle of the follow-up period.

```r
# Fleming-Harrington test statistic
fit_sex <- ten(surv_object ~ sex, data = lung_cc)
comp(fit_sex)
```

```
##                      Q         Var       Z pNorm
## 1             -13.2826     29.2085 -2.4577     5
## n           -1424.0000 303810.1223 -2.5835     4
## sqrtN        -133.2089   2591.7008 -2.6166     1
## S1             -9.0659     12.2163 -2.5938     2
## S2             -8.9740     11.9825 -2.5925     3
## FH_p=1_q=1     -2.1759      1.0197 -2.1548     6
##                maxAbsZ        Var      Q pSupBr
## 1          1.3806e+01 2.9209e+01 2.5545      5
## n          1.4410e+03 3.0381e+05 2.6143      4
## sqrtN      1.3534e+02 2.5917e+03 2.6585      1
## S1         9.1529e+00 1.2216e+01 2.6187      2
## S2         9.0560e+00 1.1982e+01 2.6162      3
## FH_p=1_q=1 2.2505e+00 1.0197e+00 2.2287      6
```

```r
lrt_mat <- attr(fit_sex, "lrt")
data.frame(
test = c("Log-Rank", "Wilcoxon", "Tarone", "Peto", "Modified-Peto", "FH(1, 1)"),
Z_squared = lrt_mat[, "Z"]^2
)
```

```
##            test        Z
## 1      Log-Rank 6.040270
## 2      Wilcoxon 6.674485
## 3        Tarone 6.846710
## 4          Peto 6.727883
## 5 Modified-Peto 6.720914
## 6      FH(1, 1) 4.643096
```

The final summary table (Z_squared) shows that regardless of the specific test used, the result is always highly significant (all Z^2 values correspond to p-values well below 0.05).

ECOG performance status (unstratified)

```r
# --- Fit the Kaplan-Meier model for ph_ecog ---
km_ecog_fit <- survfit(surv_object ~ ph_ecog, data = lung_cc)

summary(km_ecog_fit)
```

```
## Call: survfit(formula = surv_object ~ ph_ecog, data = lung_cc)
##
##                 ph_ecog=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5     47       1    0.979  0.0210       0.9383        1.000
##    15     46       1    0.957  0.0294       0.9014        1.000
##    31     45       1    0.936  0.0357       0.8688        1.000
##    53     44       1    0.915  0.0407       0.8385        0.998
##    81     43       1    0.894  0.0450       0.8097        0.986
##   147     42       1    0.872  0.0487       0.7820        0.973
##   176     40       1    0.851  0.0521       0.7543        0.959
##   246     34       1    0.826  0.0563       0.7223        0.944
```

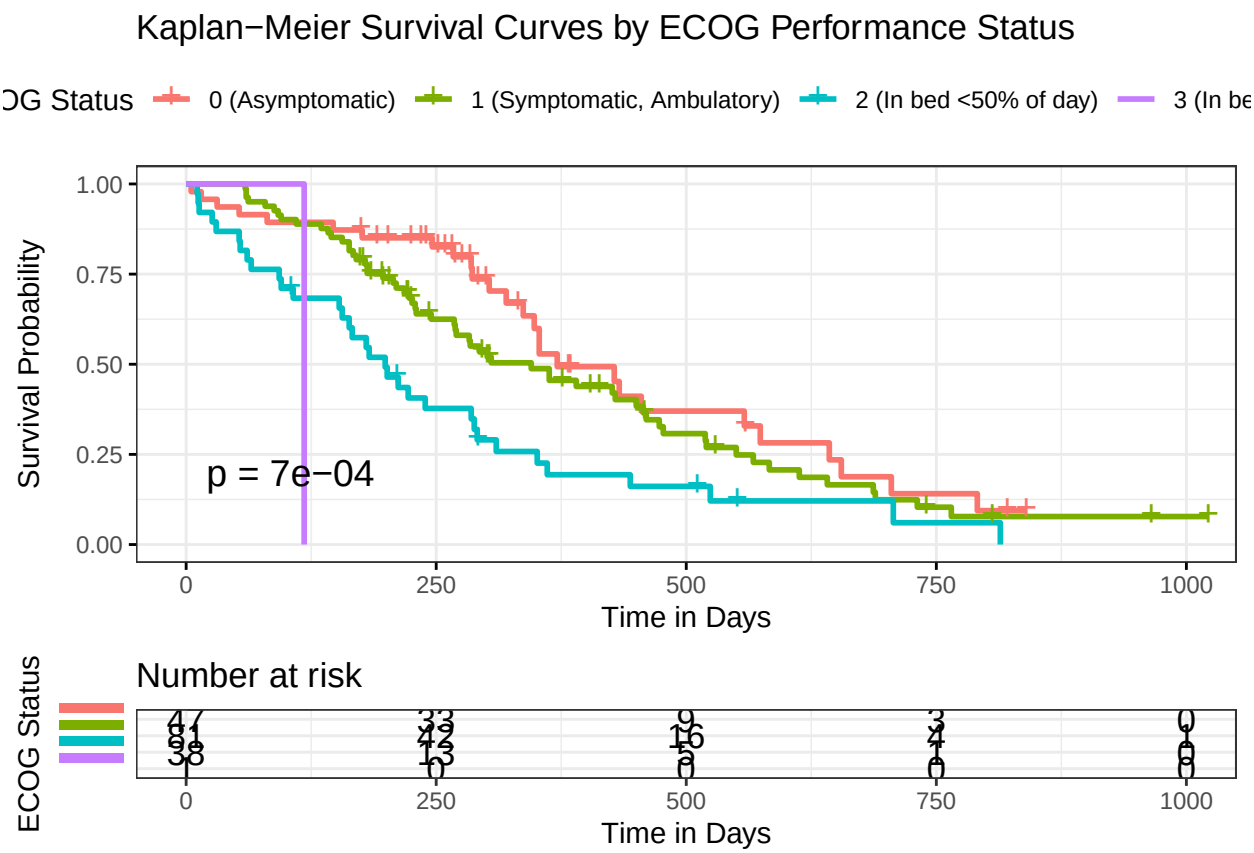
```
##   267     30       1    0.798  0.0607       0.6874        0.926
##   285     26       1    0.767  0.0657       0.6487        0.908
##   286     25       1    0.737  0.0699       0.6116        0.887
##   303     22       1    0.703  0.0743       0.5716        0.865
##   320     21       1    0.670  0.0779       0.5331        0.841
##   337     19       1    0.634  0.0814       0.4933        0.816
##   348     18       1    0.599  0.0842       0.4549        0.789
##   353     17       2    0.529  0.0878       0.3818        0.732
##   371     15       1    0.493  0.0887       0.3468        0.702
##   428     12       1    0.452  0.0904       0.3057        0.669
##   433     11       1    0.411  0.0910       0.2664        0.635
##   455     10       1    0.370  0.0907       0.2289        0.598
##   558      9       1    0.329  0.0895       0.1930        0.561
##   574      7       1    0.282  0.0882       0.1527        0.520
##   643      6       1    0.235  0.0851       0.1155        0.478
##   655      5       1    0.188  0.0800       0.0816        0.433
##   705      4       1    0.141  0.0725       0.0515        0.386
##   791      3       1    0.094  0.0617       0.0260        0.340
##
##                ph_ecog=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    59     81       1   0.9877  0.0123       0.9639        1.000
##    60     80       2   0.9630  0.0210       0.9227        1.000
##    62     78       1   0.9506  0.0241       0.9046        0.999
##    79     77       1   0.9383  0.0267       0.8873        0.992
##    88     76       1   0.9259  0.0291       0.8706        0.985
##    92     75       1   0.9136  0.0312       0.8544        0.977
##    95     74       1   0.9012  0.0331       0.8385        0.969
##   110     73       1   0.8889  0.0349       0.8230        0.960
##   135     72       1   0.8765  0.0366       0.8078        0.951
##   142     71       1   0.8642  0.0381       0.7927        0.942
##   145     70       1   0.8519  0.0395       0.7779        0.933
##   156     69       1   0.8395  0.0408       0.7633        0.923
##   163     68       2   0.8148  0.0432       0.7345        0.904
##   167     66       1   0.8025  0.0442       0.7203        0.894
##   170     65       1   0.7901  0.0452       0.7062        0.884
##   179     62       1   0.7774  0.0463       0.6918        0.874
##   181     61       2   0.7519  0.0481       0.6632        0.852
##   197     57       1   0.7387  0.0491       0.6485        0.841
##   207     54       1   0.7250  0.0500       0.6333        0.830
##   210     53       1   0.7113  0.0509       0.6182        0.818
##   218     52       1   0.6977  0.0517       0.6033        0.807
##   223     49       1   0.6834  0.0526       0.5877        0.795
##   226     47       1   0.6689  0.0535       0.5719        0.782
##   229     46       1   0.6543  0.0542       0.5562        0.770
##   230     45       1   0.6398  0.0550       0.5407        0.757
##   245     43       1   0.6249  0.0557       0.5248        0.744
##   268     42       1   0.6100  0.0563       0.5091        0.731
##   269     41       1   0.5952  0.0568       0.4936        0.718
##   270     40       1   0.5803  0.0573       0.4781        0.704
##   283     39       1   0.5654  0.0578       0.4628        0.691
##   284     38       1   0.5505  0.0581       0.4476        0.677
##   293     37       1   0.5356  0.0584       0.4325        0.663
##   301     35       1   0.5203  0.0587       0.4171        0.649
```

```
##   305     32       1   0.5041  0.0591       0.4006        0.634
##   345     31       1   0.4878  0.0594       0.3843        0.619
##   363     30       2   0.4553  0.0597       0.3521        0.589
##   390     27       1   0.4384  0.0598       0.3355        0.573
##   426     24       1   0.4202  0.0601       0.3175        0.556
##   429     23       1   0.4019  0.0602       0.2997        0.539
##   450     22       1   0.3836  0.0601       0.2821        0.522
##   457     21       1   0.3654  0.0600       0.2648        0.504
##   460     19       1   0.3461  0.0598       0.2467        0.486
##   473     18       1   0.3269  0.0595       0.2288        0.467
##   477     17       1   0.3077  0.0590       0.2112        0.448
##   519     16       1   0.2884  0.0584       0.1940        0.429
##   520     15       1   0.2692  0.0576       0.1770        0.409
##   550     13       1   0.2485  0.0568       0.1588        0.389
##   567     12       1   0.2278  0.0557       0.1411        0.368
##   583     11       1   0.2071  0.0543       0.1238        0.346
##   613     10       1   0.1864  0.0527       0.1071        0.324
##   641      9       1   0.1657  0.0507       0.0909        0.302
##   687      8       1   0.1450  0.0484       0.0753        0.279
##   689      7       1   0.1243  0.0457       0.0604        0.256
##   731      6       1   0.1035  0.0425       0.0463        0.232
##   765      4       1   0.0777  0.0390       0.0290        0.208
##
##                ph_ecog=2
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    11     38       1   0.9737  0.0260       0.9241        1.000
##    12     37       1   0.9474  0.0362       0.8790        1.000
##    13     36       1   0.9211  0.0437       0.8392        1.000
##    26     35       1   0.8947  0.0498       0.8023        0.998
##    30     34       1   0.8684  0.0548       0.7673        0.983
##    53     33       1   0.8421  0.0592       0.7338        0.966
##    54     32       1   0.8158  0.0629       0.7014        0.949
##    61     31       1   0.7895  0.0661       0.6699        0.930
##    65     30       1   0.7632  0.0690       0.6393        0.911
##    93     29       1   0.7368  0.0714       0.6093        0.891
##    95     28       1   0.7105  0.0736       0.5800        0.870
##   107     26       1   0.6832  0.0756       0.5499        0.849
##   153     25       1   0.6559  0.0774       0.5204        0.827
##   156     24       1   0.6285  0.0789       0.4915        0.804
##   163     23       1   0.6012  0.0800       0.4632        0.780
##   166     22       1   0.5739  0.0809       0.4353        0.757
##   180     21       1   0.5466  0.0815       0.4080        0.732
##   183     20       1   0.5192  0.0819       0.3811        0.707
##   199     19       1   0.4919  0.0820       0.3547        0.682
##   201     18       1   0.4646  0.0819       0.3288        0.656
##   212     16       1   0.4355  0.0818       0.3014        0.629
##   222     15       1   0.4065  0.0813       0.2747        0.602
##   239     14       1   0.3775  0.0805       0.2485        0.573
##   285     13       1   0.3484  0.0794       0.2229        0.545
##   288     12       1   0.3194  0.0779       0.1980        0.515
##   291     11       1   0.2904  0.0760       0.1738        0.485
##   310      9       1   0.2581  0.0741       0.1470        0.453
##   351      8       1   0.2258  0.0715       0.1214        0.420
##   361      7       1   0.1936  0.0682       0.0970        0.386
```

```
##   444      6       1   0.1613  0.0640       0.0741        0.351
##   524      4       1   0.1210  0.0594       0.0462        0.317
##   707      2       1   0.0605  0.0521       0.0112        0.327
##   814      1       1   0.0000     NaN           NA           NA
##
##                 ph_ecog=3
##         time       n.risk      n.event     survival      std.err lower 95% CI
##          118            1            1            0          NaN           NA
## upper 95% CI
##           NA
```

```r
ggsurvplot(
  km_ecog_fit,
  data = lung_cc,

  # --- Add Key Statistical Information ---
  pval = TRUE,                # Display the log-rank test p-value
  conf.int = FALSE,            # Show 95% confidence intervals

  # --- Add Contextual Information ---
  risk.table = TRUE,          # Include a table showing the number of patients at risk
  risk.table.y.text.col = TRUE,# Color the risk table text to match curves
  risk.table.y.text = FALSE,   # Remove y-axis tick labels from the risk table

  # --- Customize Aesthetics and Labels ---
  title = "Kaplan-Meier Survival Curves by ECOG Performance Status",
  xlab = "Time in Days",
  ylab = "Survival Probability",
  legend.title = "ECOG Status",
  # Provide clear, descriptive labels for each group in the legend
  legend.labs = c(
    "0 (Asymptomatic)",
    "1 (Symptomatic, Ambulatory)",
    "2 (In bed <50% of day)",
    "3 (In bed >50% of day)"
  ),
  ggtheme = theme_bw()         # Use a clean black and white theme
)
```

```
## Ignoring unknown labels:
## * colour : "ECOG Status"
```




```
# --- Perform the test ---
log_rank_ecog <- survdiff(surv_object ~ ph_ecog, data = lung_cc, rho = 0)
print(log_rank_ecog)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog, data = lung_cc, rho = 0)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47       27   38.054     3.211     4.737
## ph_ecog=1 81       59   62.376     0.183     0.383
## ph_ecog=2 38       33   19.394     9.546    11.483
## ph_ecog=3  1        1    0.176     3.864     3.895
##
##  Chisq= 17  on 3 degrees of freedom, p= 7e-04
```

```
wilcoxon_ecog <- survdiff(surv_object ~ ph_ecog, data = lung_cc, rho = 1)
print(wilcoxon_ecog)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog, data = lung_cc, rho = 1)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47   14.231   22.518     3.050     6.299
## ph_ecog=1 81   33.801   37.054     0.286     0.826
## ph_ecog=2 38   23.250   12.392     9.513    15.668
## ph_ecog=3  1    0.844    0.162     2.878     3.144
##
##  Chisq= 20.8  on 3 degrees of freedom, p= 1e-04
```

The fact that the Wilcoxon test yields a smaller p-value than the log-rank test suggests that the survival differences between the groups are particularly pronounced early in the follow-up period.

ECOG performance status (stratified by sex)

```
log_rank_ecog_bysex <- survdiff(surv_object ~ ph_ecog + strata(sex), data = lung_cc, rho = 0)
print(log_rank_ecog_bysex)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog + strata(sex), data = lung_cc,
##     rho = 0)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47       27   38.330     3.349     5.003
## ph_ecog=1 81       59   62.336     0.178     0.381
## ph_ecog=2 38       33   19.131    10.053    12.287
## ph_ecog=3  1        1    0.203     3.136     3.173
##
##  Chisq= 17.1  on 3 degrees of freedom, p= 7e-04
```

```
wilcoxon_ecog_bysex <- survdiff(surv_object ~ ph_ecog + strata(sex), data = lung_cc, rho = 1)
print(wilcoxon_ecog_bysex)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog + strata(sex), data = lung_cc,
##     rho = 1)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47   13.478   22.324     3.505      7.26
## ph_ecog=1 81   34.002   37.497     0.326      0.96
## ph_ecog=2 38   23.686   11.987    11.419     18.70
## ph_ecog=3  1    0.825    0.184     2.226      2.47
##
##  Chisq= 23.4  on 3 degrees of freedom, p= 3e-05
```

Log-rank: The result is virtually identical to the unstratified test. The Chi-square statistic and the p-value have barely changed. It means that the strong predictive power of ECOG status is independent of the patient's sex. The effect is not being confounded by sex.

Wilcoxon: After stratifying by sex, the Chi-square value increased (from 20.8 to 23.4) and the p-value became even smaller. This suggests that after controlling for the baseline survival differences between sexes, the effect of ECOG status on early deaths becomes even more pronounced and clear.

Final summary for stratified analysis: ECOG is an Independent Predictor: The effect of ECOG performance status on survival is not explained away by the patient's sex. It is a strong, independent prognostic factor. Sex is not acting as a major confounder in the relationship between ECOG status and survival. Whether a patient is male or female, having a worse ECOG score (e.g., a score of 2 vs. 0) is strongly associated with a significantly poorer survival outcome. The effect holds true within both groups.

Cox model

Univariate Cox Regression (Single-variable screening)

```
# List of all predictor variables to test
predictors <- c("age", "sex", "ph_ecog", "ph_karno", "pat_karno", "meal_cal", "wt_loss")
```

```r
# Create an empty list to store results
univariate_results <- list()

# Loop through each predictor and fit a univariate Cox model
for (var in predictors) {
  formula <- as.formula(paste("Surv(time, event) ~", var))
  cox_model <- coxph(formula, data = lung_cc)

  # Extract coefficients, HR, CI, and p-value
  summary_model <- summary(cox_model)

  univariate_results[[var]] <- data.frame(
    Variable = var,
    HR = summary_model$conf.int[1, "exp(coef)"],
    Lower_CI = summary_model$conf.int[1, "lower .95"],
    Upper_CI = summary_model$conf.int[1, "upper .95"],
    P_value = summary_model$coefficients[1, "Pr(>|z|)"]
  )
}

# Combine all results into one table
univariate_table <- do.call(rbind, univariate_results)
rownames(univariate_table) <- NULL

# Display the table
print(univariate_table)
```

```
##    Variable        HR  Lower_CI   Upper_CI     P_value
## 1       age 1.0200874 0.9988278  1.0417995 0.064195884
## 2       sex 0.6192558 0.4212238  0.9103896 0.014790685
## 3   ph_ecog 4.8252501 1.2331287 18.8812725 0.023758640
## 4  ph_karno 0.9878729 0.9749832  1.0009329 0.068635439
## 5 pat_karno 0.9812708 0.9694664  0.9932189 0.002199523
## 6  meal_cal 0.9998804 0.9994063  1.0003547 0.620970110
## 7   wt_loss 1.0001532 0.9871172  1.0133615 0.981737216
```

```r
# Create a more detailed summary table using gtsummary
tbl_univariate <-
  tbl_uvregression(
    lung_cc %>% select(time, event, age, sex, ph_ecog, ph_karno, pat_karno, meal_cal, wt_loss),
    method = coxph,
    y = Surv(time, event),
    exponentiate = TRUE,
    pvalue_fun = ~style_pvalue(.x, digits = 3)
  ) %>%
  bold_p(t = 0.05) %>%
  bold_labels()

tbl_univariate
```

In univariable Cox models, female sex was associated with a substantially lower hazard of death compared with males (HR 0.62, 95% CI 0.42–0.91, p=0.015). Worse ECOG performance status showed a significant linear trend toward higher mortality risk (p for linear trend = 0.024). Higher patient-rated Karnofsky score was strongly protective (HR 0.98 per point, 95% CI 0.97–0.99, p=0.002). Age and physician-rated Karnofsky score exhibited borderline associations with survival, whereas meal calories and weight loss were

Characteristic	N	HR	95% CI	p-value
age	167	1.02	1.00, 1.04	0.064
sex	167			
Male		—	—	
Female		0.62	0.42, 0.91	**0.015**
ph_ecog	167			
ph_ecog.L		4.83	1.23, 18.9	**0.024**
ph_ecog.Q		1.63	0.58, 4.57	0.356
ph_ecog.C		1.08	0.64, 1.85	0.765
ph_karno	167	0.99	0.97, 1.00	0.069
pat_karno	167	0.98	0.97, 0.99	**0.002**
meal_cal	167	1.00	1.00, 1.00	0.621
wt_loss	167	1.00	0.99, 1.01	0.982

Abbreviations: CI = Confidence Interval, HR = Hazard Ratio

not significantly associated with overall survival in univariable analyses

Variable Selection

```
significant_vars <- univariate_table %>%
  filter(P_value < 0.10) %>%
  pull(Variable)

print("Variables selected for multivariable model (p < 0.10):")
```

```
## [1] "Variables selected for multivariable model (p < 0.10):"
```

```
print(significant_vars)
```

```
## [1] "age"       "sex"       "ph_ecog"   "ph_karno"  "pat_karno"
```

Covariates with p < 0.10 in univariable Cox models were considered as candidates for the multivariable model. Given the strong clinical relevance, age and sex were retained a priori. Because physician- and patient-rated Karnofsky scores are highly correlated, we included only the patient-rated score (pat_karno), which showed a stronger univariable association (p=0.002), in the primary model.

Multivariable Cox Regression (Initial Model)

```
cox_multi_model <- coxph(Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
                         data = lung_cc)

summary(cox_multi_model)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
##     data = lung_cc)
##
##   n= 167, number of events= 120
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age        0.006488  1.006509  0.011276  0.575   0.5650
```

```
## sexFemale -0.483159  0.616832  0.198121 -2.439   0.0147 *
## ph_ecog.L  1.325682  3.764752  0.711628  1.863   0.0625 .
## ph_ecog.Q  0.417938  1.518827  0.531404  0.786   0.4316
## ph_ecog.C  0.119441  1.126866  0.285060  0.419   0.6752
## pat_karno -0.006584  0.993438  0.008366 -0.787   0.4313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0065     0.9935    0.9845    1.0290
## sexFemale    0.6168     1.6212    0.4183    0.9095
## ph_ecog.L    3.7648     0.2656    0.9332   15.1871
## ph_ecog.Q    1.5188     0.6584    0.5360    4.3037
## ph_ecog.C    1.1269     0.8874    0.6445    1.9702
## pat_karno    0.9934     1.0066    0.9773    1.0099
##
## Concordance= 0.653  (se = 0.031 )
## Likelihood ratio test= 21.4  on 6 df,   p=0.002
## Wald test            = 22.47  on 6 df,   p=0.001
## Score (logrank) test = 24.29  on 6 df,   p=5e-04
```

In the initial multivariable Cox model including age, sex, ECOG performance status, and patient-rated Karnofsky score, the overall model was statistically significant (likelihood ratio test p = 0.002), with a concordance index of 0.65, indicating moderate discriminative ability. After adjustment for the other covariates, female patients had a substantially lower hazard of death compared with males (HR 0.62, 95% CI 0.42–0.91, p = 0.015). ECOG performance status showed a borderline linear trend toward higher mortality (HR 3.76 for a one-step worsening in ECOG, 95% CI 0.93–15.19, p = 0.063). In contrast, age and patient-rated Karnofsky score were not significantly associated with survival in this multivariable model (all p > 0.40).

Tied Events Handling + Model Simplification

```
cox_multi_efron <- coxph(
  Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
  data = lung_cc,
  ties = "efron"
)

cox_multi_breslow <- coxph(
  Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
  data = lung_cc,
  ties = "breslow"
)

summary(cox_multi_efron)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
##     data = lung_cc, ties = "efron")
##
##   n= 167, number of events= 120
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age         0.006488  1.006509  0.011276  0.575   0.5650
## sexFemale -0.483159  0.616832  0.198121 -2.439   0.0147 *
```

```
## ph_ecog.L  1.325682  3.764752  0.711628  1.863   0.0625 .
## ph_ecog.Q  0.417938  1.518827  0.531404  0.786   0.4316
## ph_ecog.C  0.119441  1.126866  0.285060  0.419   0.6752
## pat_karno -0.006584  0.993438  0.008366 -0.787   0.4313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0065     0.9935    0.9845    1.0290
## sexFemale    0.6168     1.6212    0.4183    0.9095
## ph_ecog.L    3.7648     0.2656    0.9332   15.1871
## ph_ecog.Q    1.5188     0.6584    0.5360    4.3037
## ph_ecog.C    1.1269     0.8874    0.6445    1.9702
## pat_karno    0.9934     1.0066    0.9773    1.0099
##
## Concordance= 0.653  (se = 0.031 )
## Likelihood ratio test= 21.4  on 6 df,   p=0.002
## Wald test            = 22.47  on 6 df,   p=0.001
## Score (logrank) test = 24.29  on 6 df,   p=5e-04
```

```
summary(cox_multi_breslow)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ age + sex + ph_ecog + pat_karno,
##     data = lung_cc, ties = "breslow")
##
##   n= 167, number of events= 120
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age         0.006475  1.006496  0.011274  0.574   0.5657
## sexFemale -0.482427  0.617284  0.198125 -2.435   0.0149 *
## ph_ecog.L  1.326123  3.766413  0.711627  1.864   0.0624 .
## ph_ecog.Q  0.418666  1.519933  0.531395  0.788   0.4308
## ph_ecog.C  0.119903  1.127388  0.285042  0.421   0.6740
## pat_karno -0.006594  0.993427  0.008361 -0.789   0.4303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0065     0.9935    0.9845    1.0290
## sexFemale    0.6173     1.6200    0.4186    0.9102
## ph_ecog.L    3.7664     0.2655    0.9337   15.1938
## ph_ecog.Q    1.5199     0.6579    0.5364    4.3067
## ph_ecog.C    1.1274     0.8870    0.6448    1.9711
## pat_karno    0.9934     1.0066    0.9773    1.0098
##
## Concordance= 0.653  (se = 0.031 )
## Likelihood ratio test= 21.37  on 6 df,   p=0.002
## Wald test            = 22.45  on 6 df,   p=0.001
## Score (logrank) test = 24.26  on 6 df,   p=5e-04
```

To assess the impact of tied event times, we refit the multivariable Cox model using both the Efron and Breslow methods for handling ties. The estimated hazard ratios, confidence intervals, and p-values were nearly identical across the two approaches, and the concordance index remained 0.65 in both models. These findings suggest that ties had minimal influence on the results, and we therefore used the Efron method in

all subsequent analyses.

Model Simplification

```
# Remove non-significant variables (age, pat_karno)
# Test 3 simplified models

# Model A: Sex only (most parsimonious)
cox_model_A <- coxph(Surv(time, event) ~ sex, data = lung_cc)

# Model B: Sex + ph_ecog (categorical)
cox_model_B <- coxph(Surv(time, event) ~ sex + ph_ecog, data = lung_cc)

# Model C: Sex + ph_ecog (as numeric/ordinal)
lung_cc <- lung_cc %>%
  mutate(ph_ecog_num = as.numeric(as.character(ph_ecog)))

cox_model_C <- coxph(Surv(time, event) ~ sex + ph_ecog_num, data = lung_cc)

# Display all models
cat("========================================\n")
```

```
## ========================================
```

```
cat("Model A: Sex Only\n")
```

```
## Model A: Sex Only
```

```
cat("========================================\n")
```

```
## ========================================
```

```
summary(cox_model_A)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ sex, data = lung_cc)
##
##   n= 167, number of events= 120
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## sexFemale -0.4792    0.6193   0.1966 -2.437   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sexFemale    0.6193      1.615    0.4212    0.9104
##
## Concordance= 0.567  (se = 0.025 )
## Likelihood ratio test= 6.25  on 1 df,   p=0.01
## Wald test            = 5.94  on 1 df,   p=0.01
## Score (logrank) test = 6.05  on 1 df,   p=0.01
```

```
cat("\n========================================\n")
```

```
##
## ========================================
```

```r
cat("Model B: Sex + ECOG (Categorical)\n")
```

```
## Model B: Sex + ECOG (Categorical)
```

```r
cat("========================================\n")
```

```
## ========================================
```

```r
summary(cox_model_B)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ sex + ph_ecog, data = lung_cc)
##
##   n= 167, number of events= 120
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## sexFemale -0.49977   0.60667  0.19741 -2.532   0.0114 *
## ph_ecog.L  1.47356   4.36475  0.69684  2.115   0.0345 *
## ph_ecog.Q  0.37911   1.46098  0.52802  0.718   0.4728
## ph_ecog.C  0.04551   1.04657  0.27299  0.167   0.8676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sexFemale    0.6067     1.6483    0.4120    0.8933
## ph_ecog.L    4.3647     0.2291    1.1138   17.1045
## ph_ecog.Q    1.4610     0.6845    0.5190    4.1124
## ph_ecog.C    1.0466     0.9555    0.6129    1.7870
##
## Concordance= 0.646  (se = 0.03 )
## Likelihood ratio test= 20.39  on 4 df,   p=4e-04
## Wald test            = 21.86  on 4 df,   p=2e-04
## Score (logrank) test = 23.52  on 4 df,   p=1e-04
```

```r
cat("\n========================================\n")
```

```
##
## ========================================
```

```r
cat("Model C: Sex + ECOG (Numeric)\n")
```

```
## Model C: Sex + ECOG (Numeric)
```

```r
cat("========================================\n")
```

```
## ========================================
```

```r
summary(cox_model_C)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ sex + ph_ecog_num, data = lung_cc)
##
##   n= 167, number of events= 120
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## sexFemale   -0.5101    0.6004   0.1969 -2.591 0.009579 **
## ph_ecog_num  0.4825    1.6201   0.1323  3.647 0.000266 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## sexFemale      0.6004     1.6655    0.4082    0.8832
## ph_ecog_num    1.6201     0.6172    1.2501    2.0998
##
## Concordance= 0.641  (se = 0.031 )
## Likelihood ratio test= 19.48  on 2 df,   p=6e-05
## Wald test            = 19.35  on 2 df,   p=6e-05
## Score (logrank) test = 19.62  on 2 df,   p=5e-05
```

```
# Model comparison
cat("\n========================================\n")
```

```
##
## ========================================
```

```
cat("MODEL COMPARISON SUMMARY\n")
```

```
## MODEL COMPARISON SUMMARY
```

```
cat("========================================\n")
```

```
## ========================================
```

```
# Extract metrics
aic_A <- AIC(cox_model_A)
aic_B <- AIC(cox_model_B)
aic_C <- AIC(cox_model_C)

c_A <- summary(cox_model_A)$concordance[1]
c_B <- summary(cox_model_B)$concordance[1]
c_C <- summary(cox_model_C)$concordance[1]

cat("\nModel A (sex only):\n")
```

```
##
## Model A (sex only):
```

```
cat("  AIC =", round(aic_A, 2), "\n")
```

```
##   AIC = 1011.99
```

```
cat("  C-index =", round(c_A, 3), "\n")
```

```
##   C-index = 0.567
```

```
cat("\nModel B (sex + ph_ecog categorical):\n")
```

```
##
## Model B (sex + ph_ecog categorical):
```

```
cat("  AIC =", round(aic_B, 2), "\n")
```

```
##   AIC = 1003.84
```

```
cat("  C-index =", round(c_B, 3), "\n")
```

```
##   C-index = 0.646
```

```
cat("\nModel C (sex + ph_ecog numeric):\n")
```

```
##
## Model C (sex + ph_ecog numeric):
cat("  AIC =", round(aic_C, 2), "\n")

##   AIC = 1000.75
cat("  C-index =", round(c_C, 3), "\n")

##   C-index = 0.641
cat("\nNote: Lower AIC = Better fit\n")

##
## Note: Lower AIC = Better fit
cat("      Higher C-index = Better discrimination\n")

##       Higher C-index = Better discrimination
# Likelihood Ratio Tests
cat("\n========================================\n")

##
## ========================================
cat("LIKELIHOOD RATIO TESTS\n")

## LIKELIHOOD RATIO TESTS
cat("========================================\n")

## ========================================
cat("\n--- Test 1: Model A vs Model B ---\n")

##
## --- Test 1: Model A vs Model B ---
cat("(Does adding ph_ecog categorical improve the model?)\n")

## (Does adding ph_ecog categorical improve the model?)
lrt_AB <- anova(cox_model_A, cox_model_B, test = "Chisq")
print(lrt_AB)

## Analysis of Deviance Table
##  Cox model: response is  Surv(time, event)
##  Model 1: ~ sex
##  Model 2: ~ sex + ph_ecog
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -504.99
## 2 -497.92 14.144  3   0.002715 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
cat("\n--- Test 2: Model A vs Model C ---\n")

##
## --- Test 2: Model A vs Model C ---
cat("(Does adding ph_ecog numeric improve the model?)\n")

## (Does adding ph_ecog numeric improve the model?)
```

```r
lrt_AC <- anova(cox_model_A, cox_model_C, test = "Chisq")
print(lrt_AC)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, event)
##  Model 1: ~ sex
##  Model 2: ~ sex + ph_ecog_num
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -504.99
## 2 -498.38 13.235  1  0.0002747 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat("\n--- Test 3: Model B vs Model C ---\n")
```

```
##
## --- Test 3: Model B vs Model C ---
```

```r
cat("(Is categorical or numeric ph_ecog better?)\n")
```

```
## (Is categorical or numeric ph_ecog better?)
```

```r
lrt_BC <- anova(cox_model_C, cox_model_B, test = "Chisq")
print(lrt_BC)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, event)
##  Model 1: ~ sex + ph_ecog_num
##  Model 2: ~ sex + ph_ecog
##    loglik  Chisq Df Pr(>|Chi|)
## 1 -498.38
## 2 -497.92 0.9089  2     0.6348
```

Starting from a multivariable model including age, sex, ECOG performance status, and patient-rated Karnofsky score, we removed age (p=0.565) and patient-rated Karnofsky score (p=0.431) as they were not statistically significant after adjustment for other covariates. We then compared three reduced Cox models: (A) sex only, (B) sex plus ECOG treated as a categorical factor, and (C) sex plus ECOG treated as a numeric ordinal variable (0–3).

Adding ECOG status substantially improved model fit compared with the sex-only model. For model B, the likelihood ratio test versus model A yielded χ^2=14.14 on 3 df (p=0.003), with concordance increasing from 0.567 to 0.646. For model C, the likelihood ratio test versus model A yielded χ^2=13.24 on 1 df (p<0.001), with concordance of 0.641. Models B and C demonstrated similar performance (AIC=1003.84 vs 1000.75; C-index=0.646 vs 0.641), with no significant difference between them (χ^2=0.91 on 2 df, p=0.635).

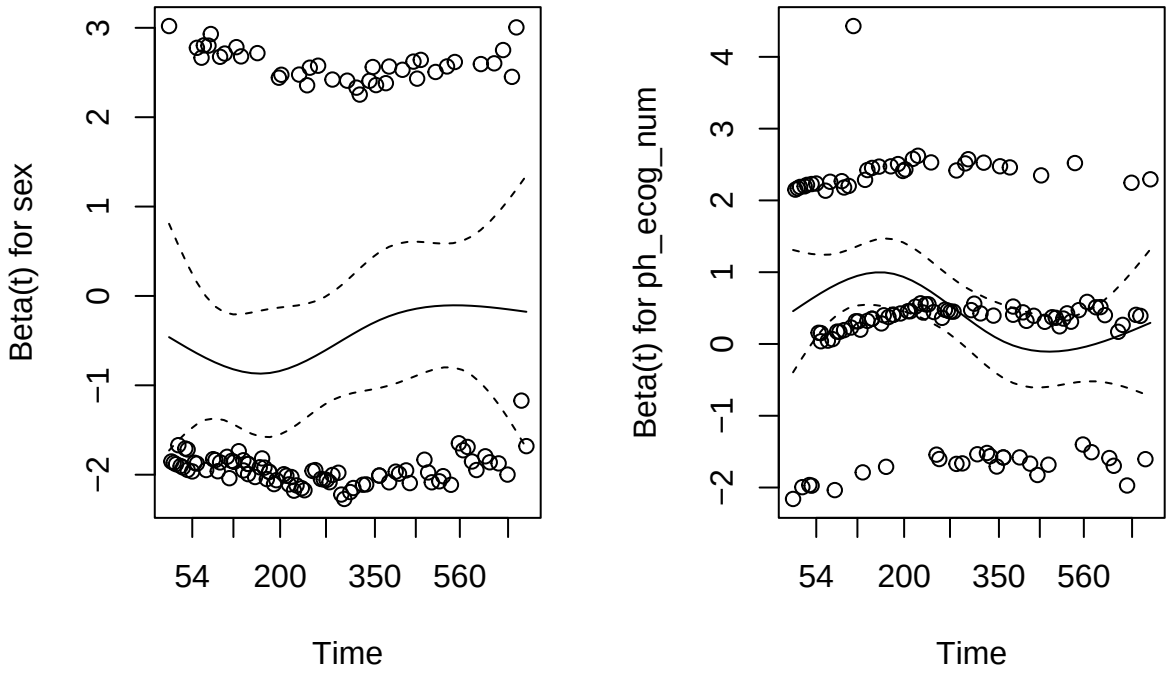
We selected model C (sex + ECOG as a numeric ordinal variable) as the final Cox model based on the principle of parsimony: it achieved similar predictive performance with fewer parameters (2 vs 4) and offered a simpler, more interpretable linear effect for ECOG performance status.

Check assumptions

```r
ph_test = cox.zph(cox_model_C)
ph_test
```

```
##             chisq df     p
## sex          1.24  1 0.266
## ph_ecog_num  5.27  1 0.022
## GLOBAL       6.19  2 0.045
```

```r
par(mfrow = c(1,2))
plot(ph_test)
```




```r
par(mfrow = c(1,1))
```

The proportional hazards assumption was evaluated using Schoenfeld residuals. The effect of sex met the PH assumption (p = 0.266), with residual plots showing no meaningful time trend. However, ph_ecog_num showed a statistically significant deviation from proportional hazards (p = 0.022), and its residual plot displayed a clear time-varying pattern. The global test was also statistically significant (p = 0.045), further indicating model-level violation of the PH assumption. These findings suggest that the effect of ECOG performance status may vary over time, and the current Cox model may not fully satisfy the PH assumption. Thus, we might considering remove the ECOG variable, and only include sex into our final survival model.

AFT Model Analysis (Addressing PH Violation)

Since `ph_ecog` violated the proportional hazards assumption (p = 0.022), we fit Accelerated Failure Time (AFT) models as an alternative approach. AFT models do not require the PH assumption and provide **time ratios** that are often more clinically interpretable.

```r
# ====================================
# Variable Selection for AFT Models (using Weibull as reference)
# ====================================
# We use Weibull for initial variable selection because it's most commonly used

# Full model
aft_full_weibull <- survreg(Surv(time, event) ~ sex + age + ph_ecog_num +
                            ph_karno + pat_karno + meal_cal + wt_loss,
                            data = lung_cc, dist = "weibull")
summary(aft_full_weibull)
```

```
##
## Call:
## survreg(formula = Surv(time, event) ~ sex + age + ph_ecog_num +
```

```
##     ph_karno + pat_karno + meal_cal + wt_loss, data = lung_cc,
##     dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept)  7.41e+00   1.06e+00  6.99 2.7e-12
## sexFemale    3.92e-01   1.42e-01  2.76 0.00577
## age         -6.50e-03   7.99e-03 -0.81 0.41613
## ph_ecog_num -5.28e-01   1.55e-01 -3.41 0.00065
## ph_karno    -1.64e-02   7.57e-03 -2.17 0.03012
## pat_karno    8.40e-03   5.60e-03  1.50 0.13389
## meal_cal    -9.05e-06   1.79e-04 -0.05 0.95974
## wt_loss      9.37e-03   5.34e-03  1.75 0.07933
## Log(scale)  -3.58e-01   7.24e-02 -4.94 7.8e-07
##
## Scale= 0.699
##
## Weibull distribution
## Loglik(model)= -826.9   Loglik(intercept only)= -841.1
##  Chisq= 28.32 on 7 degrees of freedom, p= 0.00019
## Number of Newton-Raphson Iterations: 5
## n= 167
```

```
# Check collinearity
cor(lung_cc[, c("ph_ecog_num", "ph_karno")], use = "complete.obs")
```

```
##             ph_ecog_num   ph_karno
## ph_ecog_num   1.0000000 -0.8226974
## ph_karno     -0.8226974  1.0000000
```

```
# Compare models
aft_with_karno <- survreg(Surv(time, event) ~ sex + ph_ecog_num + ph_karno,
                          data = lung_cc, dist = "weibull")
aft_reduced <- survreg(Surv(time, event) ~ sex + ph_ecog_num,
                       data = lung_cc, dist = "weibull")

# LRT and AIC
anova(aft_reduced, aft_with_karno)
```

```
##                           Terms Resid. Df    -2*LL      Test Df Deviance
## 1            sex + ph_ecog_num       163 1662.563           NA       NA
## 2 sex + ph_ecog_num + ph_karno       162 1659.296 +ph_karno  1  3.26724
##     Pr(>Chi)
## 1         NA
## 2 0.07067645
```

Although ph_karno was significant in the full model (p=0.030), the likelihood ratio test shows that adding it to a model already containing sex and ph_ecog does not significantly improve model fit (p=0.071 > 0.05). This, combined with the high collinearity between ph_ecog and ph_karno (r=-0.82) and minimal AIC difference (Δ=1.27), indicates that ph_karno does not provide meaningful additional information beyond what ph_ecog already captures. Therefore, we select the more parsimonious reduced model with only sex and ph_ecog_num as predictors. This decision is further supported by the fact that ph_ecog has a stronger effect (p=0.001 vs p=0.030) and is more widely used in clinical practice.

Fit Multiple AFT Models

```
# Fit AFT models with different distributions
# Model: time ~ sex + ph_ecog_num
```

```r
# 1. Exponential
aft_exp <- survreg(Surv(time, event) ~ sex + ph_ecog_num,
                   data = lung_cc, dist = "exponential")

# 2. Weibull (most common, allows PH interpretation)
aft_weibull <- survreg(Surv(time, event) ~ sex + ph_ecog_num,
                       data = lung_cc, dist = "weibull")

# 3. Log-Normal
aft_lnorm <- survreg(Surv(time, event) ~ sex + ph_ecog_num,
                     data = lung_cc, dist = "lognormal")

# 4. Log-Logistic
aft_llogis <- survreg(Surv(time, event) ~ sex + ph_ecog_num,
                      data = lung_cc, dist = "loglogistic")
```

Model Comparison

```r
# Compare models using AIC
model_comparison <- data.frame(
  Model = c("Exponential", "Weibull", "Log-Normal", "Log-Logistic"),
  LogLik = c(logLik(aft_exp)[1], logLik(aft_weibull)[1],
             logLik(aft_lnorm)[1], logLik(aft_llogis)[1]),
  AIC = c(AIC(aft_exp), AIC(aft_weibull),
          AIC(aft_lnorm), AIC(aft_llogis)),
  Parameters = c(3, 4, 4, 4)
) %>%
  mutate(Delta_AIC = AIC - min(AIC)) %>%
  arrange(AIC)

kable(model_comparison, digits = 2,
      caption = "AFT Model Comparison (Lower AIC = Better)") %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Table 1: AFT Model Comparison (Lower AIC = Better)

Model	LogLik	AIC	Parameters	Delta_AIC
Weibull	-831.28	1670.56	4	0.00
Log-Logistic	-834.33	1676.65	4	6.09
Exponential	-839.73	1685.46	3	14.90
Log-Normal	-843.30	1694.59	4	24.03

```r
best_model <- model_comparison$Model[1]
cat("\nBest model by AIC:", best_model, "\n")
```

```
##
## Best model by AIC: Weibull
```

Likelihood Ratio Test

```r
# Test if Weibull significantly better than Exponential
lrt_stat <- 2 * (logLik(aft_weibull)[1] - logLik(aft_exp)[1])
```

```r
p_value <- pchisq(lrt_stat, df = 1, lower.tail = FALSE)

cat("LR Chi-square:", round(lrt_stat, 3), "\n")
```

```
## LR Chi-square: 16.9
```

```r
cat("df: 1\n")
```

```
## df: 1
```

```r
cat("p-value:", format.pval(p_value, digits = 3), "\n\n")
```

```
## p-value: 3.94e-05
```

```r
if (p_value < 0.05) {
  cat("Result: Weibull is significantly better (p < 0.05)\n")
} else {
  cat("Result: No significant difference\n")
}
```

```
## Result: Weibull is significantly better (p < 0.05)
```

Final AFT Model Results

```r
cat("================================================\n")
```

```
## ================================================
```

```r
cat("FINAL WEIBULL AFT MODEL\n")
```

```
## FINAL WEIBULL AFT MODEL
```

```r
cat("================================================\n\n")
```

```
## ================================================
```

```r
cat("Model Selection Rationale:\n")
```

```
## Model Selection Rationale:
```

```r
cat("- Weibull had the lowest AIC among all distributions tested\n")
```

```
## - Weibull had the lowest AIC among all distributions tested
```

```r
cat("- LR test: Weibull significantly better than Exponential (p < 0.001)\n")
```

```
## - LR test: Weibull significantly better than Exponential (p < 0.001)
```

```r
cat("- Variables: sex + ph_ecog_num (reduced model)\n\n")
```

```
## - Variables: sex + ph_ecog_num (reduced model)
```

```r
cat("Weibull Parameters:\n")
```

```
## Weibull Parameters:
```

```r
cat("- Scale ($\\sigma$):", round(aft_weibull$scale, 4), "\n")
```

```
## - Scale ($\sigma$): 0.7263
```

```r
cat("- Shape ($\\kappa = 1/\\sigma$):", round(1/aft_weibull$scale, 4), "\n")
```

```
## - Shape ($\kappa = 1/\sigma$): 1.3768
```

```r
cat("- Interpretation: $\\kappa > 1$ indicates INCREASING hazard over time\n\n")
```

```
## - Interpretation: $\kappa > 1$ indicates INCREASING hazard over time
```

```r
# Get coefficients
aft_sum <- summary(aft_weibull)$table

coefs <- aft_sum[-1, "Value"]
se    <- aft_sum[-1, "Std. Error"]
names(coefs) <- rownames(aft_sum)[-1]

n_coef <- length(coefs)

final_aft_results <- data.frame(
  Variable   = names(coefs),
  Beta_AFT   = as.numeric(coefs),
  SE         = as.numeric(se),
  Time_Ratio = exp(coefs),
  Lower_95CI = exp(coefs - 1.96 * se),
  Upper_95CI = exp(coefs + 1.96 * se),
  p_value    = 2 * pnorm(-abs(coefs / se)),
  stringsAsFactors = FALSE
)

kable(final_aft_results, digits = 3,
      caption = "Weibull AFT Model Results")
```

Table 2: Weibull AFT Model Results

	Variable	Beta_AFT	SE	Time_Ratio	Lower_95CI	Upper_95CI	p_value
sexFemale	sexFemale	0.373	0.144	1.452	1.094	1.926	0.01
ph_ecog_num	ph_ecog_num	-0.354	0.098	0.702	0.579	0.850	0.00
Log(scale)	Log(scale)	-0.320	0.072	0.726	0.631	0.836	0.00

In the Weibull AFT model ($\sigma = 0.73$, $\kappa = 1.38$), both sex and ECOG performance status were significant independent predictors of survival time. Sex: Female patients demonstrated significantly longer median survival compared to males (Time Ratio [TR] = 1.45, 95% CI: 1.09–1.93, p = 0.01), with an estimated 45% increase in median survival time. ECOG Performance Status: Each one-unit worsening in ECOG score was associated with a 30% reduction in median survival time (TR = 0.70, 95% CI: 0.58–0.85, $p < 0.001$).

Clinical Interpretation

```r
cat("============================================\n")
```

```
## ============================================
```

```r
cat("CLINICAL INTERPRETATION (AFT Model)\n")
```

```
## CLINICAL INTERPRETATION (AFT Model)
```

```r
cat("============================================\n\n")
```

```
## ============================================
```

```r
# Sex effect
tr_sex <- final_aft_results$Time_Ratio[1]
pct_sex <- (tr_sex - 1) * 100

cat("1. Sex (Female vs Male):\n")
```

```
## 1. Sex (Female vs Male):
```

```r
cat("   Time Ratio:", round(tr_sex, 3), "\n")
```

```
##    Time Ratio: 1.452
```

```r
if (tr_sex > 1) {
  cat("   Females have", round(pct_sex, 1), "% LONGER median survival\n\n")
} else {
  cat("   Females have", round(abs(pct_sex), 1), "% SHORTER median survival\n\n")
}
```

```
##    Females have 45.2 % LONGER median survival
```

```r
# ECOG effect
tr_ecog <- final_aft_results$Time_Ratio[2]
pct_ecog <- (tr_ecog - 1) * 100

cat("2. ECOG (per 1-unit increase):\n")
```

```
## 2. ECOG (per 1-unit increase):
```

```r
cat("   Time Ratio:", round(tr_ecog, 3), "\n")
```

```
##    Time Ratio: 0.702
```

```r
if (tr_ecog < 1) {
  cat("   Each 1-unit worsening reduces survival by", round(abs(pct_ecog), 1), "%\n\n")
} else {
  cat("   Each 1-unit worsening increases survival by", round(pct_ecog, 1), "%\n\n")
}
```

```
##    Each 1-unit worsening reduces survival by 29.8 %
```

```r
# Example
cat("Example:\n")
```

```
## Example:
```

```r
cat("If a male with ECOG=0 has median survival of 300 days:\n")
```

```
## If a male with ECOG=0 has median survival of 300 days:
```

```r
cat("  - Male, ECOG=1:   ", round(300 * tr_ecog, 0), "days\n")
```

```
##   - Male, ECOG=1:    211 days
```

```r
cat("  - Female, ECOG=0: ", round(300 * tr_sex, 0), "days\n")
```

```
##   - Female, ECOG=0:  435 days
```

```r
cat("  - Female, ECOG=1: ", round(300 * tr_sex * tr_ecog, 0), "days\n")
```

```
##   - Female, ECOG=1:  306 days
```

In the Weibull AFT model, female sex is associated with ~45% longer survival time compared with males, adjusting for ECOG. Each 1-point worsening in ECOG score is associated with ~30% shorter survival time. For example, if a male patient with ECOG 0 has a median survival of 300 days, the model predicts 211 days for a male with ECOG 1, 435 days for a female with ECOG 0, and 306 days for a female with ECOG 1.

Weibull AFT to Hazard Ratios

```
if (best_model == "Weibull") {
  cat("============================================\n")
  cat("WEIBULL AFT to PH CONVERSION\n")
  cat("============================================\n\n")

  sigma <- aft_weibull$scale
  beta_aft <- coef(aft_weibull)[-1]
  beta_ph <- -beta_aft / sigma
  hr <- exp(beta_ph)

  conversion_table <- data.frame(
    Variable = names(beta_aft),
    AFT_TimeRatio = exp(beta_aft),
    Converted_HR = hr,
    Cox_HR = exp(coef(cox_model_C))
  )

  kable(conversion_table, digits = 3,
        caption = "AFT vs Cox: Time Ratios and Hazard Ratios") %>%
    kable_styling(bootstrap_options = c("striped", "hover"))

  cat("\nNote: Weibull AFT and Cox PH give similar HRs\n")
}
```

```
## ============================================
## WEIBULL AFT to PH CONVERSION
## ============================================
##
##
## Note: Weibull AFT and Cox PH give similar HRs
```

Summary

```
cat("============================================\n")
```

```
## ============================================
```

```
cat("SUMMARY: AFT vs COX\n")
```

```
## SUMMARY: AFT vs COX
```

```
cat("============================================\n\n")
```

```
## ============================================
```

```
cat("1. PH ASSUMPTION:\n")
```

```
## 1. PH ASSUMPTION:
```

```
cat("   - Cox: ph_ecog violated (p=0.022)\n")
```

```
##    - Cox: ph_ecog violated (p=0.022)
```

```
cat("   - AFT: NO assumption required  \n\n")
```

```
##    - AFT: NO assumption required
```

```
cat("2. INTERPRETATION:\n")
```

```
## 2. INTERPRETATION:
```

```
cat("   - Cox: Hazard Ratios\n")
```

```
##    - Cox: Hazard Ratios
```

```
cat("   - AFT: Time Ratios (more intuitive)\n\n")
```

```
##    - AFT: Time Ratios (more intuitive)
```

```
cat("3. KEY FINDINGS:\n")
```

```
## 3. KEY FINDINGS:
```

```
cat("   - Female sex: protective effect\n")
```

```
##    - Female sex: protective effect
```

```
cat("   - ECOG worse: shorter survival\n")
```

```
##    - ECOG worse: shorter survival
```

```
cat("   - Consistent across methods  \n\n")
```

```
##    - Consistent across methods
```