Survival_Analysis.R

Nikita Maurya

2025-11-13

```r
# loading necessary packages for data manipulation and survival analysis
library(readxl)  # to read Excel files
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```r
library(dplyr) # for data wrangling and cleaning
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(survival) # for survival analysis functions
library(survminer) # for plotting survival curves nicely
```

```
## Warning: package 'survminer' was built under R version 4.3.3
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```r
library(ggplot2) # load ggplot2 to save plots as PNG

data = read_excel("Clinical_Data_Validation_Cohort.xlsx") # load the clinical dataset from Excel
View(data)  # view the dataset to check its structure

# renaming columns
data <- data %>% rename(
  Event = `Event (death: 1, alive: 0)`,
  Stage = `Stage (TNM 8th edition)`
)

# Convert columns to factors
```

```r
data$Grade <- as.factor(data$Grade)
data$Stage <- as.factor(data$Stage)
data$Sex <- as.factor(data$Sex)
data$Cigarette <- as.factor(data$Cigarette)
data$Type.Adjuvant <- as.factor(data$Type.Adjuvant)
data$EGFR <- as.factor(data$EGFR)
data$KRAS <- as.factor(data$KRAS)

summary(data) # quick summary to check data types and missing values
```

```
##    Patient ID        Survival time (days)      Event          Tumor size (cm)
##  Length:95          Min.   :  50.0        Min.   :0.0000    Min.   :1.100
##  Class :character   1st Qu.: 898.5        1st Qu.:0.0000    1st Qu.:2.000
##  Mode  :character   Median :1760.0        Median :0.0000    Median :2.400
##                     Mean   :1471.5        Mean   :0.4105    Mean   :2.855
##                     3rd Qu.:1981.0        3rd Qu.:1.0000    3rd Qu.:3.500
##                     Max.   :2532.0        Max.   :1.0000    Max.   :7.000
##
##  Grade       Stage          Age           Sex          Cigarette   Pack per year
##  1: 6    IB     :21    Min.   :48.00   Female:68    Current:11   Min.   :  0.000
##  2:48    IA2    :20    1st Qu.:60.00   Male  :27    Former :62   1st Qu.:  2.125
##  3:41    IA3    :20    Median :67.00                Never  :22   Median : 25.000
##          IIIA   :13    Mean   :66.59                             Mean   : 29.254
##          IIB    :12    3rd Qu.:72.50                             3rd Qu.: 45.000
##          IIIB   : 3    Max.   :88.00                             Max.   :105.000
##          (Other): 6
##   Type.Adjuvant      batch                              EGFR           KRAS
##  Chemo   :17    Min.   :1.000    Negative            :66   Negative:32
##  Chemorad: 4    1st Qu.:2.000    NA                  : 9   NA      :30
##  NA      : 2    Median :3.000    Exon 19             : 8   G12C    : 9
##  None    :71    Mean   :2.432    Exon 21 (L858R)     : 4   G12V    : 8
##  XRT     : 1    3rd Qu.:3.000    Exon 21             : 3   G12D    : 6
##                 Max.   :3.000    Exon 19 (15bp delete): 2  G12A    : 3
##                                  (Other)             : 3   (Other) : 7
```

```r
#### Survival Analysis ######

s = Surv(data$`Survival time (days)`, data$Event) # create a survival object using time and event colum
s # :1 → event (death) happened and # + → censored (alive or lost to follow-up).
```

```
##  [1] 2329  2532+ 2271+ 2193+ 2387+ 2225+ 2240+ 2314+  299  2295+ 2135+ 1956+
## [13] 2278+ 1927   837   453  2238+ 2248+ 1435  1922+ 1300  2318+ 2059+ 1865+
## [25]  978+ 1961+ 2240+ 2186+ 2041  1810  2237   647   874  2021+ 2256+   66
## [37]  855   631   592  1435+ 1760   228+ 1079  2022+ 2059+ 1614+ 1904+  895
## [49] 1986+ 1546+  135+  990+ 1976+  942  1373  1161   354   676  1579  1886+
## [61] 1834+  811  1057  1937+  681   579   208  1860+ 1862+  768+ 1841+  366
## [73] 1349  1834  1054  1353   923   902+ 1868+ 1430+   50+ 1217   335+  403+
## [85] 1945+  819  1881+ 1207+ 1889+ 1033  1919+ 1911+ 1935+ 1760+ 1922+
```

```r
sfit = survfit(Surv(`Survival time (days)`, Event) ~ 1, data = data) # baseline hazard
sfit1 = survfit(Surv(`Survival time (days)`, Event) ~ Sex, data = data) # effect of sex
sfit2 = survfit(Surv(`Survival time (days)`, Event) ~ Grade, data = data) # effect of grade
sfit3 = survfit(Surv(`Survival time (days)`, Event) ~ Type.Adjuvant, data = data) # effect of therapy

# summarize the survival fits
```

```r
summary(sfit)
```

```
## Call: survfit(formula = Surv(`Survival time (days)`, Event) ~ 1, data = data)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     66     94       1    0.989  0.0106        0.969        1.000
##    208     92       1    0.979  0.0150        0.950        1.000
##    299     90       1    0.968  0.0183        0.932        1.000
##    354     88       1    0.957  0.0212        0.916        0.999
##    366     87       1    0.946  0.0236        0.901        0.993
##    453     85       1    0.935  0.0258        0.885        0.987
##    579     84       1    0.923  0.0278        0.871        0.980
##    592     83       1    0.912  0.0296        0.856        0.972
##    631     82       1    0.901  0.0313        0.842        0.965
##    647     81       1    0.890  0.0328        0.828        0.957
##    676     80       1    0.879  0.0342        0.814        0.949
##    681     79       1    0.868  0.0356        0.801        0.940
##    811     77       1    0.857  0.0368        0.787        0.932
##    819     76       1    0.845  0.0380        0.774        0.923
##    837     75       1    0.834  0.0392        0.761        0.914
##    855     74       1    0.823  0.0402        0.748        0.906
##    874     73       1    0.812  0.0412        0.735        0.896
##    895     72       1    0.800  0.0422        0.722        0.887
##    923     70       1    0.789  0.0431        0.709        0.878
##    942     69       1    0.777  0.0439        0.696        0.868
##   1033     66       1    0.766  0.0448        0.683        0.859
##   1054     65       1    0.754  0.0457        0.669        0.849
##   1057     64       1    0.742  0.0464        0.656        0.839
##   1079     63       1    0.730  0.0472        0.643        0.829
##   1161     62       1    0.718  0.0479        0.631        0.819
##   1217     60       1    0.707  0.0485        0.617        0.808
##   1300     59       1    0.695  0.0492        0.605        0.798
##   1349     58       1    0.683  0.0498        0.592        0.787
##   1353     57       1    0.671  0.0503        0.579        0.777
##   1373     56       1    0.659  0.0508        0.566        0.766
##   1435     54       1    0.646  0.0513        0.553        0.755
##   1579     51       1    0.634  0.0519        0.540        0.744
##   1760     49       1    0.621  0.0524        0.526        0.732
##   1810     47       1    0.608  0.0529        0.512        0.721
##   1834     46       1    0.594  0.0534        0.498        0.709
##   1927     31       1    0.575  0.0550        0.477        0.694
##   2041     21       1    0.548  0.0588        0.444        0.676
##   2237     14       1    0.509  0.0664        0.394        0.657
##   2329      3       1    0.339  0.1453        0.146        0.786
```

```r
summary(sfit1)
```

```
## Call: survfit(formula = Surv(`Survival time (days)`, Event) ~ Sex,
##     data = data)
##
##                 Sex=Female
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     66     68       1    0.985  0.0146        0.957        1.000
##    299     65       1    0.970  0.0208        0.930        1.000
```

```
##   592    62       1    0.954  0.0257       0.905         1.000
##   631    61       1    0.939  0.0297       0.882         0.999
##   647    60       1    0.923  0.0330       0.861         0.990
##   676    59       1    0.908  0.0360       0.840         0.981
##   681    58       1    0.892  0.0386       0.819         0.971
##   811    56       1    0.876  0.0411       0.799         0.960
##   819    55       1    0.860  0.0433       0.779         0.949
##   837    54       1    0.844  0.0453       0.760         0.938
##   855    53       1    0.828  0.0472       0.741         0.926
##   874    52       1    0.812  0.0489       0.722         0.914
##   895    51       1    0.796  0.0505       0.703         0.902
##   923    49       1    0.780  0.0520       0.685         0.889
##  1033    46       1    0.763  0.0536       0.665         0.876
##  1054    45       1    0.746  0.0550       0.646         0.862
##  1057    44       1    0.729  0.0563       0.627         0.848
##  1079    43       1    0.712  0.0575       0.608         0.834
##  1161    42       1    0.695  0.0586       0.589         0.820
##  1300    40       1    0.678  0.0596       0.571         0.805
##  1349    39       1    0.661  0.0606       0.552         0.791
##  1353    38       1    0.643  0.0614       0.533         0.776
##  1579    34       1    0.624  0.0625       0.513         0.759
##  1927    23       1    0.597  0.0654       0.482         0.740
##  2041    15       1    0.557  0.0721       0.432         0.718
##
##                 Sex=Male
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   208    26       1    0.962  0.0377       0.890         1.000
##   354    25       1    0.923  0.0523       0.826         1.000
##   366    24       1    0.885  0.0627       0.770         1.000
##   453    23       1    0.846  0.0708       0.718         0.997
##   579    22       1    0.808  0.0773       0.670         0.974
##   942    21       1    0.769  0.0826       0.623         0.949
##  1217    20       1    0.731  0.0870       0.579         0.923
##  1373    19       1    0.692  0.0905       0.536         0.895
##  1435    18       1    0.654  0.0933       0.494         0.865
##  1760    17       1    0.615  0.0954       0.454         0.834
##  1810    16       1    0.577  0.0969       0.415         0.802
##  1834    15       1    0.538  0.0978       0.377         0.769
##  2237     4       1    0.404  0.1377       0.207         0.788
##  2329     1       1    0.000    NaN          NA            NA
```

```r
summary(sfit2)
```

```
## Call: survfit(formula = Surv(`Survival time (days)`, Event) ~ Grade,
##     data = data)
##
##                 Grade=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   299     6       1    0.833   0.152       0.583            1
##   895     5       1    0.667   0.192       0.379            1
##
##                 Grade=2
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    66    48       1    0.979  0.0206       0.940         1.000
##   453    45       1    0.957  0.0295       0.901         1.000
```

```
##    631     44      1    0.936  0.0360        0.868            1.000
##    647     43      1    0.914  0.0412        0.837            0.998
##    819     42      1    0.892  0.0456        0.807            0.986
##    855     41      1    0.870  0.0494        0.779            0.973
##   1079     40      1    0.849  0.0527        0.751            0.959
##   1161     39      1    0.827  0.0557        0.725            0.944
##   1300     38      1    0.805  0.0583        0.699            0.928
##   1353     37      1    0.783  0.0607        0.673            0.912
##   1435     35      1    0.761  0.0629        0.647            0.895
##   1834     31      1    0.736  0.0655        0.619            0.877
##   1927     20      1    0.700  0.0718        0.572            0.856
##   2041     15      1    0.653  0.0808        0.512            0.832
##   2237      9      1    0.580  0.0992        0.415            0.811
##
##                   Grade=3
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    208     39      1    0.974  0.0253       0.9260            1.000
##    354     37      1    0.948  0.0358       0.8804            1.000
##    366     36      1    0.922  0.0434       0.8404            1.000
##    579     35      1    0.895  0.0495       0.8034            0.998
##    592     34      1    0.869  0.0546       0.7683            0.983
##    676     33      1    0.843  0.0590       0.7347            0.967
##    681     32      1    0.816  0.0627       0.7022            0.949
##    811     30      1    0.789  0.0663       0.6694            0.930
##    837     29      1    0.762  0.0694       0.6374            0.911
##    874     28      1    0.735  0.0720       0.6063            0.890
##    923     26      1    0.706  0.0746       0.5744            0.869
##    942     25      1    0.678  0.0768       0.5433            0.847
##   1033     22      1    0.647  0.0792       0.5093            0.823
##   1054     21      1    0.617  0.0812       0.4762            0.798
##   1057     20      1    0.586  0.0828       0.4440            0.773
##   1217     18      1    0.553  0.0844       0.4103            0.746
##   1349     17      1    0.521  0.0854       0.3774            0.718
##   1373     16      1    0.488  0.0861       0.3455            0.690
##   1579     15      1    0.456  0.0863       0.3143            0.660
##   1760     14      1    0.423  0.0860       0.2840            0.630
##   1810     13      1    0.390  0.0853       0.2544            0.599
##   2329      2      1    0.195  0.1445       0.0458            0.833
```

```
summary(sfit3)
```

```
## Call: survfit(formula = Surv(`Survival time (days)`, Event) ~ Type.Adjuvant,
##     data = data)
##
##                 Type.Adjuvant=Chemo
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     66     17      1    0.941  0.0571        0.836            1.000
##    354     16      1    0.882  0.0781        0.742            1.000
##    453     15      1    0.824  0.0925        0.661            1.000
##    647     14      1    0.765  0.1029        0.587            0.995
##   1300     13      1    0.706  0.1105        0.519            0.959
##   1349     12      1    0.647  0.1159        0.455            0.919
##   1373     11      1    0.588  0.1194        0.395            0.876
##
##                 Type.Adjuvant=Chemorad
```

```
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    676      4       1     0.75   0.217        0.4259            1
##    811      3       1     0.50   0.250        0.1877            1
##    942      2       1     0.25   0.217        0.0458            1
##   1054      1       1     0.00     NaN            NA           NA
##
##                   Type.Adjuvant=NA
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    299      2       1      0.5   0.354         0.125            1
##   1435      1       1      0.0     NaN            NA           NA
##
##                   Type.Adjuvant=None
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    208     69       1    0.986  0.0144         0.958        1.000
##    366     66       1    0.971  0.0205         0.931        1.000
##    579     64       1    0.955  0.0252         0.907        1.000
##    592     63       1    0.940  0.0290         0.885        0.999
##    631     62       1    0.925  0.0322         0.864        0.990
##    819     60       1    0.910  0.0352         0.843        0.981
##    837     59       1    0.894  0.0378         0.823        0.972
##    855     58       1    0.879  0.0402         0.803        0.961
##    874     57       1    0.863  0.0423         0.784        0.951
##    895     56       1    0.848  0.0443         0.765        0.939
##    923     54       1    0.832  0.0462         0.747        0.928
##   1033     51       1    0.816  0.0481         0.727        0.916
##   1057     50       1    0.800  0.0498         0.708        0.903
##   1079     49       1    0.783  0.0514         0.689        0.891
##   1161     48       1    0.767  0.0529         0.670        0.878
##   1217     46       1    0.750  0.0543         0.651        0.865
##   1353     45       1    0.734  0.0556         0.632        0.851
##   1579     42       1    0.716  0.0569         0.613        0.837
##   1760     40       1    0.698  0.0582         0.593        0.822
##   1810     38       1    0.680  0.0595         0.573        0.807
##   1834     37       1    0.662  0.0607         0.553        0.792
##   1927     24       1    0.634  0.0641         0.520        0.773
##   2041     18       1    0.599  0.0696         0.477        0.752
##   2237     12       1    0.549  0.0797         0.413        0.730
##   2329      3       1    0.366  0.1585         0.157        0.855
##
##                   Type.Adjuvant=XRT
##           time      n.risk      n.event      survival      std.err lower 95% CI
##            681           1            1             0          NaN           NA
## upper 95% CI
##           NA
```

```
# plot Kaplan-Meier curves using ggsurvplot

p = ggsurvplot(
  sfit,
  data = data,
  xlab = "Days",
  ylab = "Survival Probability",
  title = "Overall Survival Curve"
)
```

```
p1 = ggsurvplot(
  sfit1,
  data = data,
  xlab = "Days",
  ylab = "Survival Probability",
  title = "Survival Curve by Sex"
)

p2 = ggsurvplot(
  sfit2,
  data = data,
  xlab = "Days",
  ylab = "Survival Probability",
  title = "Survival Curve by Grade"
)

p3 = ggsurvplot(
  sfit3,
  data = data,
  xlab = "Days",
  ylab = "Survival Probability",
  title = "Survival Curve by Therapy"
)

p
```
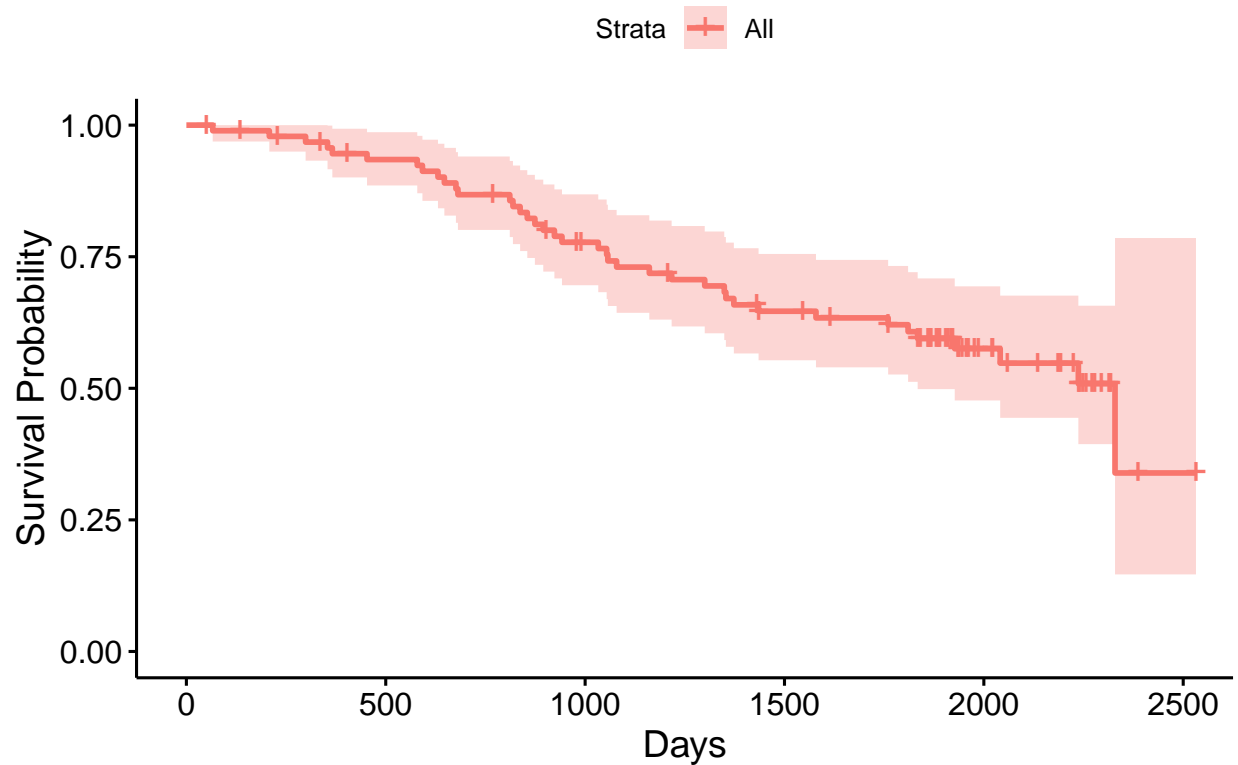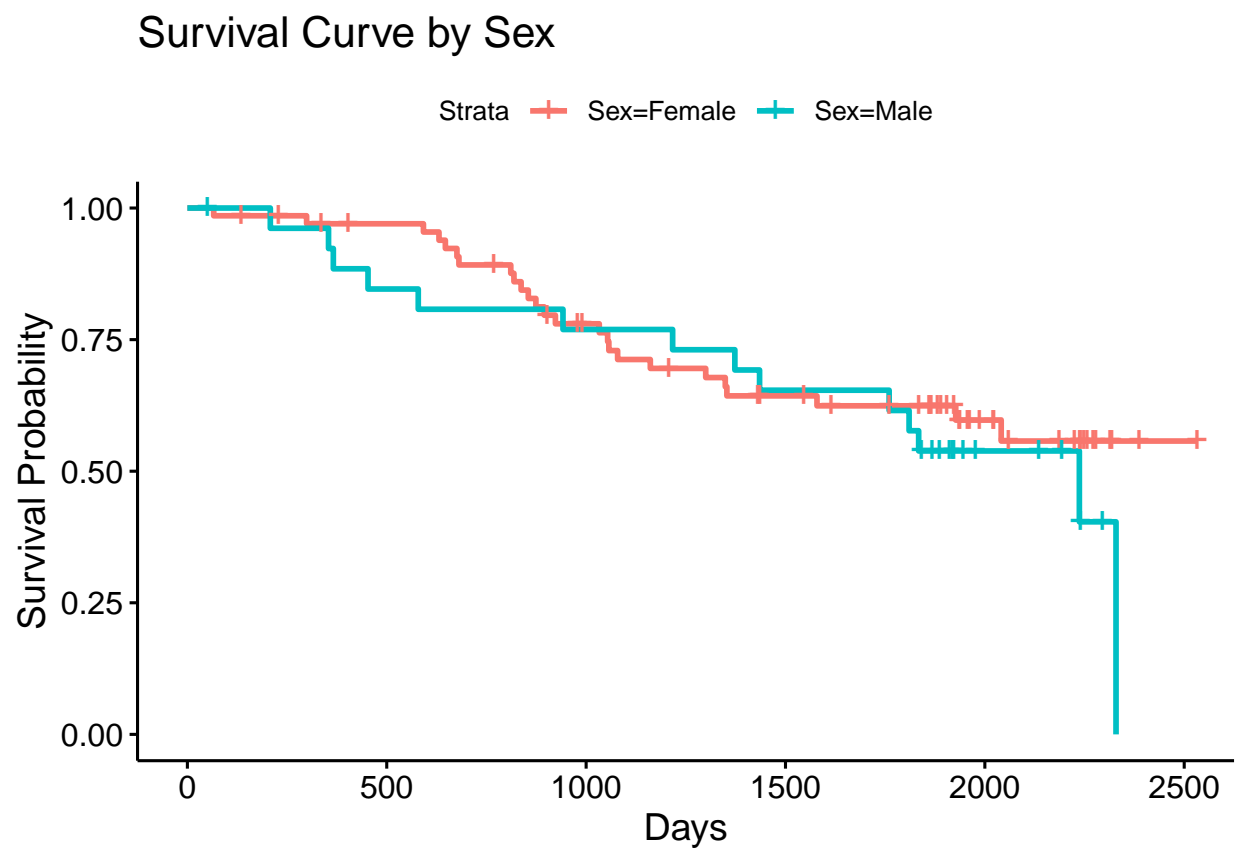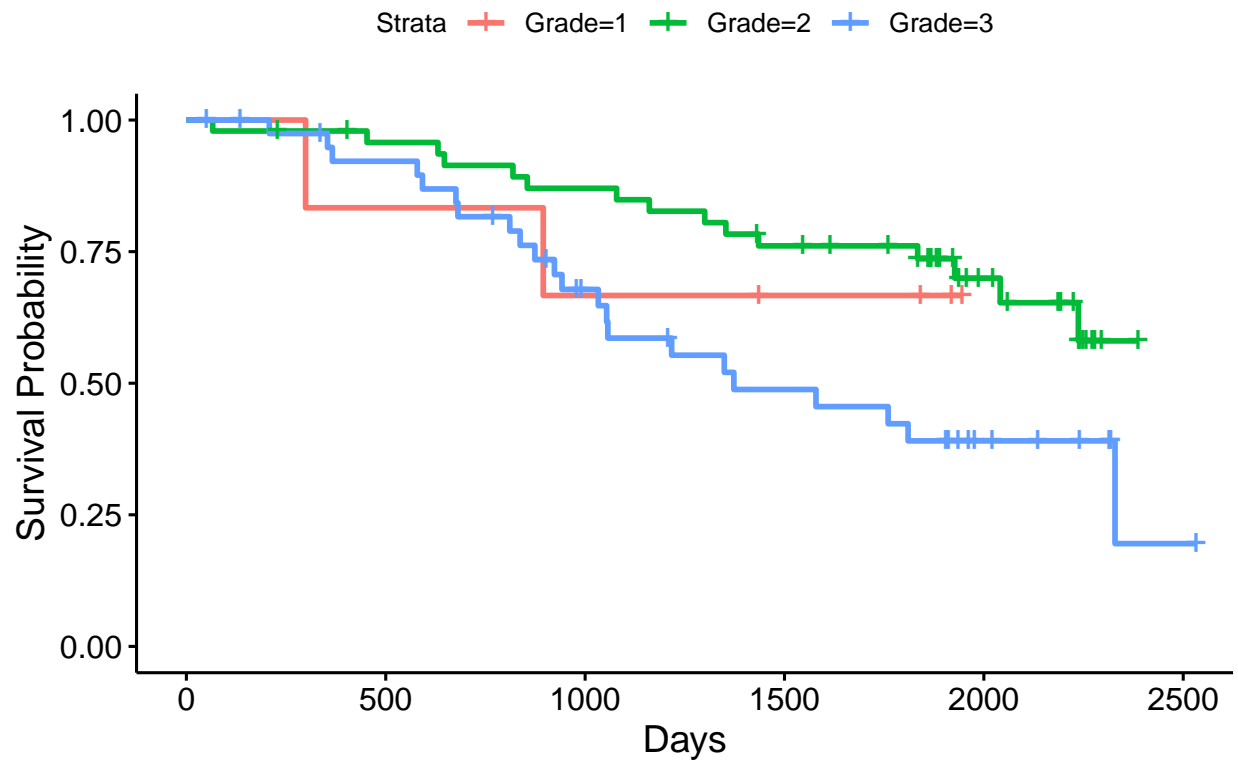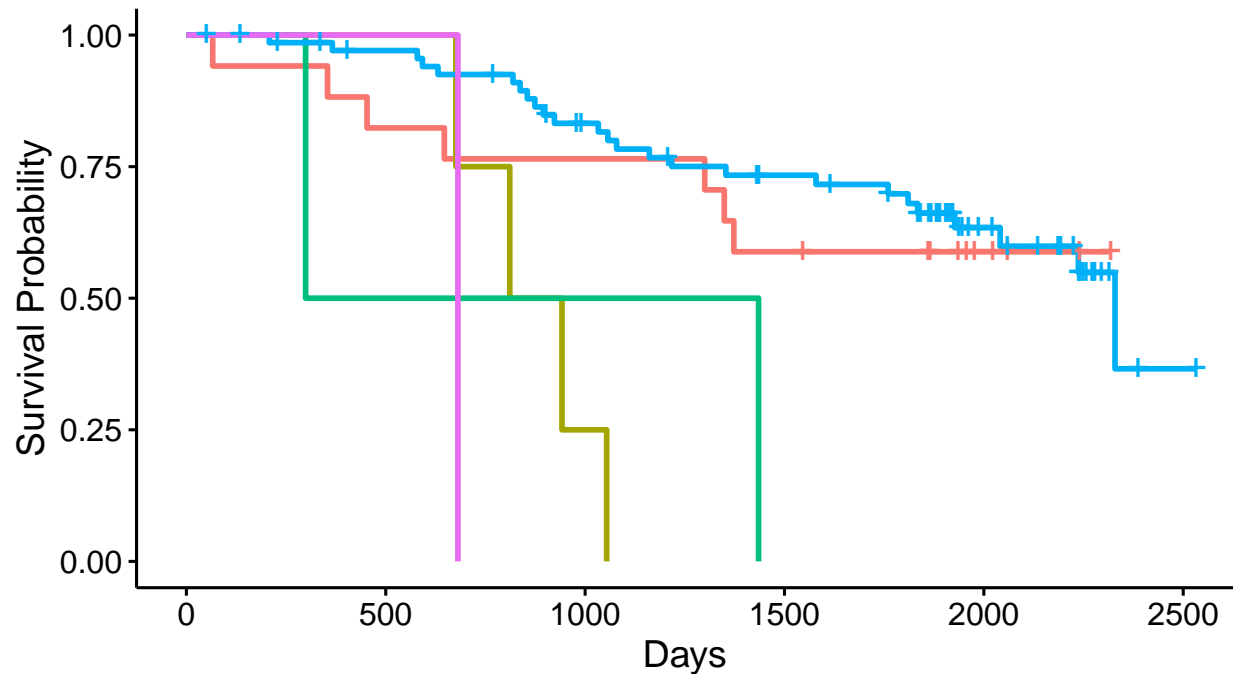

Overall Survival Curve

Survival Curve by Sex

Strata ╼ Sex=Female ╼ Sex=Male

Survival Curve by Grade

Strata ┼ Grade=1 ┼ Grade=2 ┼ Grade=3

p3

Survival Curve by Therapy


```r
# Save p
ggsave("Overall_Survival.png", plot = p$plot, width = 7, height = 5, dpi = 300)

# Save p1
ggsave("Survival_by_Sex.png", plot = p1$plot, width = 7, height = 5, dpi = 300)

# Save p2
ggsave("Survival_by_Grade.png", plot = p2$plot, width = 7, height = 5, dpi = 300)

# Save p3
ggsave("Survival_by_Therapy.png", plot = p3$plot, width = 7, height = 5, dpi = 300)

# fit Cox proportional hazards models to estimate risk (hazard ratios)

sfit_none = coxph(Surv(`Survival time (days)`, Event) ~ 1, data = data)
summary(sfit_none)
```

```
## Call:  coxph(formula = Surv(`Survival time (days)`, Event) ~ 1, data = data)
##
## Null model
##    log likelihood= -158.5004
##    n= 95
```

```r
sfit1_sex = coxph(Surv(`Survival time (days)`, Event) ~ Sex, data = data)
summary(sfit1_sex)
```

```
## Call:
```

```
## coxph(formula = Surv(`Survival time (days)`, Event) ~ Sex, data = data)
##
##   n= 95, number of events= 39
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## SexMale 0.2828    1.3268   0.3345 0.845    0.398
##
##         exp(coef) exp(-coef) lower .95 upper .95
## SexMale     1.327     0.7537    0.6888     2.556
##
## Concordance= 0.518  (se = 0.04 )
## Likelihood ratio test= 0.69  on 1 df,   p=0.4
## Wald test            = 0.71  on 1 df,   p=0.4
## Score (logrank) test = 0.72  on 1 df,   p=0.4
```

```r
sfit2_grade = coxph(Surv(`Survival time (days)`, Event) ~ Grade, data = data)
summary(sfit2_grade)
```

```
## Call:
## coxph(formula = Surv(`Survival time (days)`, Event) ~ Grade,
##     data = data)
##
##   n= 95, number of events= 39
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## Grade2 -0.2907    0.7477   0.7546 -0.385    0.700
## Grade3  0.5908    1.8054   0.7412  0.797    0.425
##
##        exp(coef) exp(-coef) lower .95 upper .95
## Grade2    0.7477     1.3374    0.1704     3.281
## Grade3    1.8054     0.5539    0.4223     7.718
##
## Concordance= 0.617  (se = 0.041 )
## Likelihood ratio test= 6.98  on 2 df,   p=0.03
## Wald test            = 6.89  on 2 df,   p=0.03
## Score (logrank) test = 7.3  on 2 df,   p=0.03
```

```r
sfit3_adjuvant = coxph(Surv(`Survival time (days)`, Event) ~ Type.Adjuvant, data = data)
summary(sfit3_adjuvant)
```

```
## Call:
## coxph(formula = Surv(`Survival time (days)`, Event) ~ Type.Adjuvant,
##     data = data)
##
##   n= 95, number of events= 39
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## Type.AdjuvantChemorad  1.7280    5.6291   0.6552  2.637  0.00835 **
## Type.AdjuvantNA        1.4968    4.4672   0.8076  1.853  0.06383 .
## Type.AdjuvantNone     -0.2100    0.8106   0.4301 -0.488  0.62535
## Type.AdjuvantXRT       2.1488    8.5748   1.1062  1.942  0.05208 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
```

```
## Type.AdjuvantChemorad    5.6291    0.1776    1.5587    20.329
## Type.AdjuvantNA          4.4672    0.2239    0.9175    21.750
## Type.AdjuvantNone        0.8106    1.2337    0.3489     1.883
## Type.AdjuvantXRT         8.5748    0.1166    0.9808    74.966
##
## Concordance= 0.605  (se = 0.04 )
## Likelihood ratio test= 13.08  on 4 df,   p=0.01
## Wald test            = 18.16  on 4 df,   p=0.001
## Score (logrank) test = 24.4  on 4 df,   p=7e-05
```