# BIOST/EPI 537 Final Group Project

Geyue Li, Cecilia Martindale, Sarah Garcia

2025-03-08

## 1. Provide an estimate of disease-free survival time for patients enrolled in this study. What are the main characteristics of this summary?

```r
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

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

```r
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

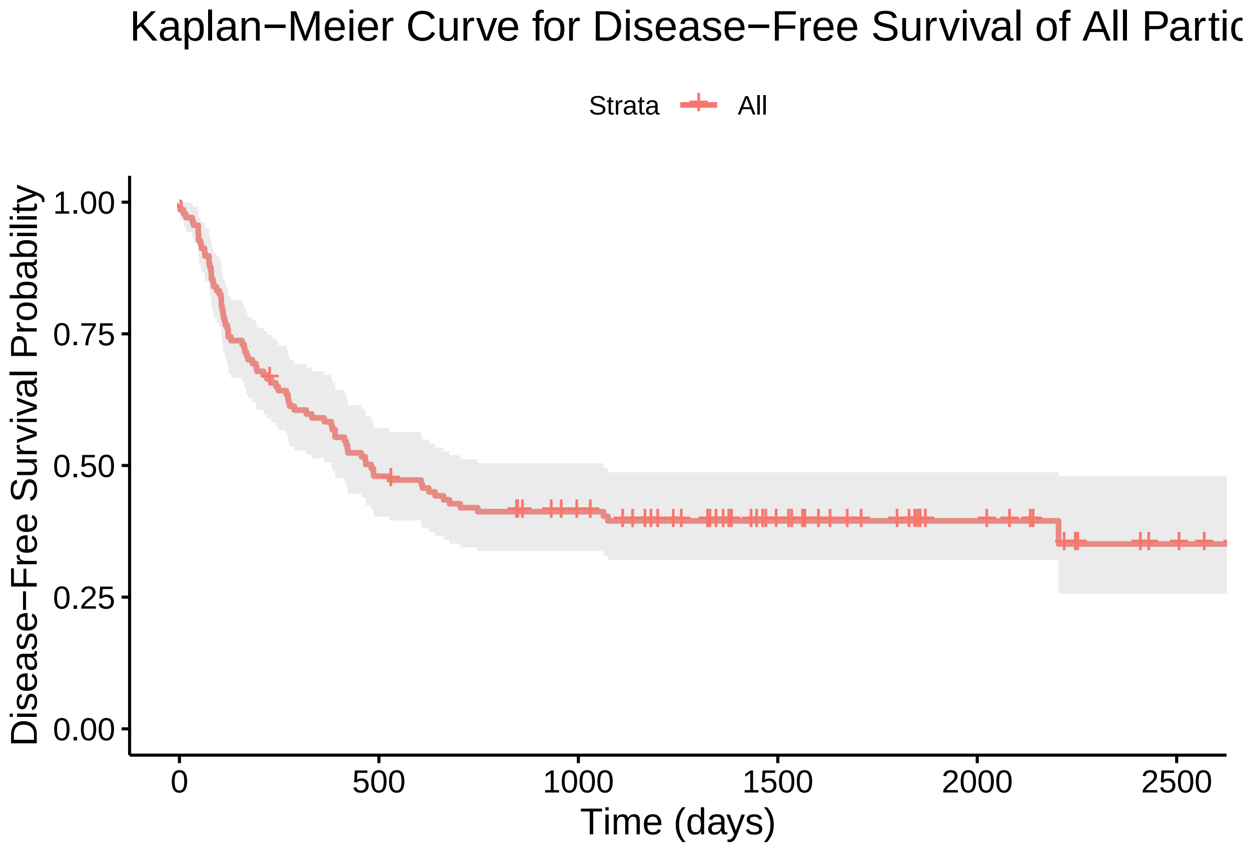
```r
library(knitr)
library(broom)
library(table1)
```

```
##
## Attaching package: 'table1'

## The following objects are masked from 'package:base':
##
##     units, units<-
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
#bmt <- read.csv("~/Documents/UW classes/survival analysis/bmt.csv")
bmt <- read.csv("~/Downloads/bmt.csv")

dfs_surv <- Surv(time = bmt$tdfs, event = bmt$deltadfs)
km_fit <- survfit(dfs_surv ~ 1)

survminer::ggsurvplot(
    fit = km_fit,
    data = bmt,
    conf.int = T,
    xlab = "Time (days)",
    ylab = "Disease-Free Survival Probability",
    title = "Kaplan-Meier Curve for Disease-Free Survival of All Participants"
)
```



```
summary(km_fit)
```

```
## Call: survfit(formula = dfs_surv ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    137       1    0.993 0.00727        0.979        1.000
##     2    136       1    0.985 0.01025        0.966        1.000
##    10    135       1    0.978 0.01250        0.954        1.000
##    16    134       1    0.971 0.01438        0.943        0.999
##    32    133       1    0.964 0.01602        0.933        0.995
##    35    132       1    0.956 0.01748        0.923        0.991
##    47    131       2    0.942 0.02003        0.903        0.982
##    48    129       2    0.927 0.02222        0.884        0.972
##    53    127       1    0.920 0.02322        0.875        0.966
##    55    126       1    0.912 0.02415        0.866        0.961
##    63    125       1    0.905 0.02504        0.857        0.956
##    64    124       1    0.898 0.02588        0.848        0.950
##    74    123       2    0.883 0.02744        0.831        0.939
##    76    121       1    0.876 0.02817        0.822        0.933
##    79    120       1    0.869 0.02886        0.814        0.927
##    80    119       2    0.854 0.03017        0.797        0.915
##    84    117       1    0.847 0.03078        0.788        0.909
##    86    116       1    0.839 0.03137        0.780        0.903
##    93    115       1    0.832 0.03193        0.772        0.897
##   100    114       1    0.825 0.03248        0.764        0.891
##   104    113       1    0.818 0.03300        0.755        0.885
##   105    112       2    0.803 0.03399        0.739        0.872
##   107    110       1    0.796 0.03445        0.731        0.866
##   109    109       1    0.788 0.03490        0.723        0.860
##   110    108       1    0.781 0.03533        0.715        0.853
##   113    107       1    0.774 0.03575        0.707        0.847
##   115    106       1    0.766 0.03615        0.699        0.841
##   120    105       1    0.759 0.03653        0.691        0.834
##   122    104       2    0.745 0.03726        0.675        0.821
##   129    102       1    0.737 0.03760        0.667        0.815
##   157    101       1    0.730 0.03793        0.659        0.808
##   162    100       1    0.723 0.03825        0.651        0.802
##   164     99       1    0.715 0.03855        0.644        0.795
##   168     98       1    0.708 0.03884        0.636        0.788
##   172     97       1    0.701 0.03912        0.628        0.782
##   183     96       1    0.693 0.03939        0.620        0.775
##   192     95       1    0.686 0.03965        0.613        0.768
##   194     94       1    0.679 0.03989        0.605        0.762
##   211     93       1    0.672 0.04013        0.597        0.755
##   219     92       1    0.664 0.04035        0.590        0.748
##   230     90       1    0.657 0.04057        0.582        0.741
##   242     89       1    0.649 0.04078        0.574        0.735
##   248     88       1    0.642 0.04098        0.567        0.728
##   268     87       1    0.635 0.04117        0.559        0.721
##   272     86       1    0.627 0.04134        0.551        0.714
##   273     85       1    0.620 0.04151        0.544        0.707
##   276     84       1    0.613 0.04167        0.536        0.700
##   288     83       1    0.605 0.04181        0.529        0.693
##   318     82       1    0.598 0.04195        0.521        0.686
##   332     81       1    0.590 0.04208        0.513        0.679
##   363     80       1    0.583 0.04219        0.506        0.672
##   381     79       1    0.576 0.04230        0.498        0.665
##   383     78       1    0.568 0.04240        0.491        0.658
##   390     77       2    0.554 0.04256        0.476        0.644
##   414     75       1    0.546 0.04263        0.469        0.636
##   418     74       1    0.539 0.04269        0.461        0.629
##   421     73       1    0.531 0.04273        0.454        0.622
##   422     72       1    0.524 0.04277        0.447        0.615
##   456     71       1    0.517 0.04280        0.439        0.608
##   466     70       1    0.509 0.04282        0.432        0.600
##   467     69       1    0.502 0.04283        0.425        0.593
##   481     68       1    0.494 0.04284        0.417        0.586
##   486     67       1    0.487 0.04283        0.410        0.579
##   487     66       1    0.480 0.04281        0.403        0.571
##   526     65       1    0.472 0.04278        0.396        0.564
##   606     63       1    0.465 0.04275        0.388        0.557
##   609     62       1    0.457 0.04272        0.381        0.549
##   625     61       1    0.450 0.04267        0.374        0.542
##   641     60       1    0.442 0.04261        0.366        0.534
##   662     59       1    0.435 0.04254        0.359        0.527
##   677     58       1    0.427 0.04247        0.352        0.519
##   704     57       1    0.420 0.04238        0.345        0.512
##   748     56       1    0.412 0.04228        0.337        0.504
##  1063     47       1    0.404 0.04228        0.329        0.496
## 1074     46       1     0.395 0.04226        0.320        0.487
## 2204      9       1     0.351 0.05587        0.257        0.479
```

## 2. How do patients in different disease groups or in different FAB classifications compare to each other with respect to other available baseline measurements?

```
baseline_vars <- c("age",  "donorage", "waittime")
cat_vars <- c("male","cmv", "donormale", "donorcmv", "mtx")
# do hospital on its own since it is 4 categories

group_summary <- bmt %>%
  group_by(disgroup) %>%
  summarise(n = n(), across(all_of(baseline_vars), list(mean = mean, sd = sd, median = median), na.rm =
            across(all_of(cat_vars), ~ mean(. == 1, na.rm = TRUE) * 100, .names = "percent_{.col}"),
            percent_Ohio=mean(hospital == 1, na.rm = TRUE) * 100,
            percent_Melbourne=mean(hospital == 2, na.rm = TRUE) * 100,
            percent_Sydney=mean(hospital == 3, na.rm = TRUE) * 100,
            percent_Philadelphia=mean(hospital == 4, na.rm = TRUE) * 100) %>%
  mutate(disgroup=case_when(disgroup==1 ~ "ALL",
                            disgroup==2 ~ "Low risk AML",
                            disgroup==3 ~ "High risk AML"))
```

```
## Warning: There was 1 warning in ‘summarise()‘.
## i In argument: ‘across(...)‘.
## i In group 1: ‘disgroup = 1‘.
## Caused by warning:
## ! The ‘...‘ argument of ‘across()‘ is deprecated as of dplyr 1.1.0.
## Supply arguments directly to ‘.fns‘ through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
kable(group_summary, caption = "Baseline Summary by Disease Group (disgroup)")
```

Table 1: Baseline Summary by Disease Group (disgroup)

| disgroup | n | age_[illegible] | [illegible] | [illegible] | [illegible] | [illegible] Philadelphia |
|---|---|---|---|---|---|---|
| ALL | 38 | 24.42[illegible] | 26.78[illegible] | 477.18[illegible] | 68.42[illegible] | [illegible]00000 |
| Low risk AML | 54 | 29.40[illegible] | 28.07[illegible] | 138.05[illegible] | 55.55[illegible] | [illegible]77778 |
| High risk AML | 45 | 30.44[illegible] | 29.93[illegible] | 268.86[illegible] | 53.33[illegible] | [illegible]33333 |

```r
# Continuous variables - ANOVA or Kruskal-Wallis
kruskal.test(age ~ disgroup, data = bmt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  age by disgroup
## Kruskal-Wallis chi-squared = 10.486, df = 2, p-value = 0.005284
```

```r
kruskal.test(donorage ~ disgroup, data = bmt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  donorage by disgroup
## Kruskal-Wallis chi-squared = 2.4723, df = 2, p-value = 0.2905
```

```r
kruskal.test(waittime ~ disgroup, data = bmt)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  waittime by disgroup
## Kruskal-Wallis chi-squared = 26.105, df = 2, p-value = 2.145e-06
```

```r
# Categorical variables - Chi-square test
table_male <- table(bmt$male, bmt$disgroup)
chisq.test(table_male)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_male
## X-squared = 2.226, df = 2, p-value = 0.3286
```

```r
table_cmv <- table(bmt$cmv, bmt$disgroup)
chisq.test(table_cmv)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_cmv
## X-squared = 3.5512, df = 2, p-value = 0.1694
```

```r
table_hospital <- table(bmt$hospital, bmt$disgroup)
fisher.test(table_hospital)
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  table_hospital
## p-value = 0.004719
## alternative hypothesis: two.sided
```

```
table_mtx <- table(bmt$mtx, bmt$disgroup)
chisq.test(table_mtx)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_mtx
## X-squared = 6.2014, df = 2, p-value = 0.04502
```

```
fab_summary <- bmt %>%
  group_by(fab) %>%
  summarise(n = n(), across(all_of(baseline_vars), list(mean = mean, sd = sd, median = median), na.rm =
            across(all_of(cat_vars), ~ mean(. == 1, na.rm = TRUE) * 100, .names = "percent_{.col}"),
            percent_Ohio=mean(hospital == 1, na.rm = TRUE) * 100,
            percent_Melbourne=mean(hospital == 2, na.rm = TRUE) * 100,
            percent_Sydney=mean(hospital == 3, na.rm = TRUE) * 100,
            percent_Philadelphia=mean(hospital == 4, na.rm = TRUE) * 100) %>%
  mutate(fab=case_when(fab==0 ~ "Otherwise",
                          fab==1 ~ "FAB Grade 4 or 5 & AML"))

kable(fab_summary, caption = "Baseline Summary by FAB Classification (fab)")
```

Table 2: Baseline Summary by FAB Classification (fab)

| fab | n | age_[illegible] | [illegible] | [illegible] | [illegible]Philadelphia |
|---|---|---|---|---|---|
| Otherwise | 92 | 28.59[illegible] | 29.00[illegible] | 308.7[illegible] | 60.86[illegible]204348 |
| FAB Grade 4 or 5 & AML | 45 | 27.88[illegible] | 26.95[illegible] | 206.3[illegible] | 53.33[illegible]00000 |

```
# Continuous variables - Wilcoxon test (since fab only has 2 levels)
wilcox.test(age ~ fab, data = bmt)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by fab
## W = 2075.5, p-value = 0.9817
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(donorage ~ fab, data = bmt)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  donorage by fab
## W = 2212, p-value = 0.5164
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(waittime ~ fab, data = bmt)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  waittime by fab
## W = 2244.5, p-value = 0.4244
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Categorical variables - Chi-square test
table_male_fab <- table(bmt$male, bmt$fab)
chisq.test(table_male_fab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_male_fab
## X-squared = 0.43028, df = 1, p-value = 0.5119
```

```r
table_cmv_fab <- table(bmt$cmv, bmt$fab)
chisq.test(table_cmv_fab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_cmv_fab
## X-squared = 0.17943, df = 1, p-value = 0.6719
```

```r
table_hospital_fab <- table(bmt$hospital, bmt$fab)
fisher.test(table_hospital_fab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_hospital_fab
## p-value = 0.2222
## alternative hypothesis: two.sided
```

```r
table_mtx_fab <- table(bmt$mtx, bmt$fab)
chisq.test(table_mtx_fab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_mtx_fab
## X-squared = 3.4444, df = 1, p-value = 0.06346
```

```
# table 1 by disease or FAB classification
bmt <- bmt %>%
  mutate(disgroupCat = factor(disgroup,
                              levels = c(1, 2, 3),
                              labels = c("All", "Low Risk AML", "High Risk AML")),
         maleCat = factor(male,
                          levels = c(0, 1),
                          labels = c("Female", "Male")),
         cmvCat = factor(cmv,
                         levels = c(0, 1),
                         labels = c("CMV Negative", "CMV Positive")),
         donormaleCat = factor(donormale,
                               levels = c(0, 1),
                               labels = c("Female Donor", "Male Donor")),
         donorcmvCat = factor(donorcmv,
                              levels = c(0, 1),
                              labels = c("CMV Negative Donor", "CMV Positive Donor")),
         fabCat = factor(fab,
                         levels = c(0, 1),
                         labels = c("Otherwise", "FAB Grade 4 or 5 and AML")),
         hospitalCat = factor(hospital,
                              levels = c(1, 2, 3, 4),
                              labels = c("OSU", "Alfred", "St. Vincent", "Hahnemann")),
         mtxCat = factor(mtx,
                         levels = c(0, 1),
                         labels = c("No", "Yes")),
         deltasCat = factor(deltas,
                            levels = c(0, 1),
                            labels = c("Alive", "Dead")),
         deltarCat = factor(deltar,
                            levels = c(0,1),
                            labels = c("Disease-free", "Relapse")),
         deltadfsCat = factor(deltadfs,
                            levels = c(0,1),
                            labels = c("Alive and Disease-free", "Dead or Relapsed")),
         deltaaCat = factor(deltaa,
                            levels = c(0,1),
                            labels = c("No aGVHD", "aGVHD")),
         deltapCat = factor(deltap,
                            levels = c(0,1),
                            labels = c("Not Recovered", "Recovered")))

label(bmt$deltarCat) <- "Replase"
label(bmt$deltasCat) <- "Death"
label(bmt$deltadfsCat) <- "Disease-free Survival"
label(bmt$deltaaCat) <- "aGVHD"
label(bmt$deltapCat) <- "Recovery of Normal Platelet Levels"
label(bmt$age) <- "Patient Age"
label(bmt$maleCat) <- "Patient Sex"
label(bmt$cmvCat) <- "Patient CMV Status"
label(bmt$donorage) <- "Donor Age"
label(bmt$donormaleCat) <- "Donor Sex"
label(bmt$donorcmvCat) <- "Donor CMV Status"
label(bmt$waittime) <- "Wait Time until Transplant (Days)"
label(bmt$fabCat) <- "Disease Subtype"
label(bmt$hospitalCat) <- "Recruitment Center"
label(bmt$mtxCat) <- "Prophylactic Methotrexate Use"

# by disease group
disgroup_tab1 <- table1(~ deltarCat + deltadfsCat + deltaaCat + deltasCat + deltapCat + age + donorage +
                        data = bmt,
                        caption = "Baseline Descriptive Statistics by Disease Group")
disgroup_tab1
```

```
# by fab classification
fab_tab1 <- table1(~ deltarCat + deltadfsCat + deltaaCat + deltasCat + deltapCat + age + donorage + wai
                   data = bmt,
                   caption = "Baseline Descriptive Statistics by FAB Classification")
fab_tab1
```

## 3. Are any of the measured baseline variables associated with differences in disease-free survival?

```
baseline_vars <- c("age", "male", "cmv", "as.factor(disgroup)",
                   "donorage", "donormale", "donorcmv",
                   "waittime", "as.factor(hospital)", "mtx")

univariate_results <- lapply(baseline_vars, function(var) {
    formula <- as.formula(paste("dfs_surv ~", var))
    cox_model <- coxph(formula, data = bmt)
    result <- tidy(cox_model, conf.int = TRUE, exponentiate=TRUE)  # Ensure confidence intervals are in
    result$variable <- var
    return(result)
})

univariate_results <- do.call(rbind, univariate_results)

colnames(univariate_results)
```

```
## [1] "term"      "estimate"  "std.error" "statistic" "p.value"   "conf.low"
## [7] "conf.high" "variable"
```

```
univariate_results <- univariate_results[, c("variable", "term", "estimate", "conf.low", "conf.high", "

kable(univariate_results, digits = 3, caption = "Univariate Cox Regression Results for Disease-Free Sur
```

Table 3: Baseline Descriptive Statistics by Disease Group

| | All | Low Risk AML | High Risk AML | Overall |
|---|---|---|---|---|
| | (N=38) | (N=54) | (N=45) | (N=137) |
| **Replase** | | | | |
| Disease-free | 26 (68.4%) | 45 (83.3%) | 24 (53.3%) | 95 (69.3%) |
| Relapse | 12 (31.6%) | 9 (16.7%) | 21 (46.7%) | 42 (30.7%) |
| **Disease-free Survival** | | | | |
| Alive and Disease-free | 14 (36.8%) | 29 (53.7%) | 11 (24.4%) | 54 (39.4%) |
| Dead or Relapsed | 24 (63.2%) | 25 (46.3%) | 34 (75.6%) | 83 (60.6%) |
| **aGVHD** | | | | |
| No aGVHD | 29 (76.3%) | 43 (79.6%) | 39 (86.7%) | 111 (81.0%) |
| aGVHD | 9 (23.7%) | 11 (20.4%) | 6 (13.3%) | 26 (19.0%) |
| **Death** | | | | |
| Alive | 14 (36.8%) | 31 (57.4%) | 11 (24.4%) | 56 (40.9%) |
| Dead | 24 (63.2%) | 23 (42.6%) | 34 (75.6%) | 81 (59.1%) |
| **Recovery of Normal Platelet Levels** | | | | |
| Not Recovered | 4 (10.5%) | 6 (11.1%) | 7 (15.6%) | 17 (12.4%) |
| Recovered | 34 (89.5%) | 48 (88.9%) | 38 (84.4%) | 120 (87.6%) |
| **Patient Age** | | | | |
| Mean (SD) | 24.4 (7.30) | 29.4 (8.76) | 30.4 (11.2) | 28.4 (9.56) |
| Median [Min, Max] | 22.5 [15.0, 42.0] | 29.5 [13.0, 50.0] | 32.0 [7.00, 52.0] | 28.0 [7.00, 52.0] |
| **Donor Age** | | | | |
| Mean (SD) | 26.8 (8.93) | 28.1 (9.24) | 29.9 (12.1) | 28.3 (10.2) |
| Median [Min, Max] | 26.0 [5.00, 48.0] | 29.5 [12.0, 54.0] | 29.0 [2.00, 56.0] | 28.0 [2.00, 56.0] |
| **Wait Time until Transplant (Days)** | | | | |
| Mean (SD) | 477 (599) | 138 (74.5) | 269 (211) | 275 (365) |
| Median [Min, Max] | 200 [74.0, 2620] | 120 [30.0, 450] | 210 [24.0, 900] | 178 [24.0, 2620] |
| **Patient Sex** | | | | |
| Female | 12 (31.6%) | 24 (44.4%) | 21 (46.7%) | 57 (41.6%) |
| Male | 26 (68.4%) | 30 (55.6%) | 24 (53.3%) | 80 (58.4%) |
| **Patient CMV Status** | | | | |
| CMV Negative | 23 (60.5%) | 28 (51.9%) | 18 (40.0%) | 69 (50.4%) |
| CMV Positive | 15 (39.5%) | 26 (48.1%) | 27 (60.0%) | 68 (49.6%) |
| **Recruitment Center** | | | | |
| OSU | 21 (55.3%) | 27 (50.0%) | 28 (62.2%) | 76 (55.5%) |
| Alfred | 8 (21.1%) | 5 (9.3%) | 4 (8.9%) | 17 (12.4%) |
| St. Vincent | 9 (23.7%) | 7 (13.0%) | 7 (15.6%) | 23 (16.8%) |
| Hahnemann | 0 (0%) | 15 (27.8%) | 6 (13.3%) | 21 (15.3%) |
| **Prophylactic Methotrexate Use** | | | | |
| No | 21 (55.3%) | 42 (77.8%) | 34 (75.6%) | 97 (70.8%) |
| Yes | 17 (44.7%) | 12 (22.2%) | 11 (24.4%) | 40 (29.2%) |

Table 4: Baseline Descriptive Statistics by FAB Classification

| | Otherwise | FAB Grade 4 or 5 and AML | Overall |
|---|---|---|---|
| | (N=92) | (N=45) | (N=137) |
| **Replase** | | | |
| Disease-free | 72 (78.3%) | 23 (51.1%) | 95 (69.3%) |
| Relapse | 20 (21.7%) | 22 (48.9%) | 42 (30.7%) |
| **Disease-free Survival** | | | |
| Alive and Disease-free | 44 (47.8%) | 10 (22.2%) | 54 (39.4%) |
| Dead or Relapsed | 48 (52.2%) | 35 (77.8%) | 83 (60.6%) |
| **aGVHD** | | | |
| No aGVHD | 74 (80.4%) | 37 (82.2%) | 111 (81.0%) |
| aGVHD | 18 (19.6%) | 8 (17.8%) | 26 (19.0%) |
| **Death** | | | |
| Alive | 45 (48.9%) | 11 (24.4%) | 56 (40.9%) |
| Dead | 47 (51.1%) | 34 (75.6%) | 81 (59.1%) |
| **Recovery of Normal Platelet Levels** | | | |
| Not Recovered | 12 (13.0%) | 5 (11.1%) | 17 (12.4%) |
| Recovered | 80 (87.0%) | 40 (88.9%) | 120 (87.6%) |
| **Patient Age** | | | |
| Mean (SD) | 28.6 (9.48) | 27.9 (9.81) | 28.4 (9.56) |
| Median [Min, Max] | 27.0 [13.0, 52.0] | 28.0 [7.00, 50.0] | 28.0 [7.00, 52.0] |
| **Donor Age** | | | |
| Mean (SD) | 29.0 (9.67) | 27.0 (11.1) | 28.3 (10.2) |
| Median [Min, Max] | 28.5 [5.00, 56.0] | 28.0 [2.00, 48.0] | 28.0 [2.00, 56.0] |
| **Wait Time until Transplant (Days)** | | | |
| Mean (SD) | 309 (427) | 206 (164) | 275 (365) |
| Median [Min, Max] | 180 [24.0, 2620] | 150 [60.0, 780] | 178 [24.0, 2620] |
| **Patient Sex** | | | |
| Female | 36 (39.1%) | 21 (46.7%) | 57 (41.6%) |
| Male | 56 (60.9%) | 24 (53.3%) | 80 (58.4%) |
| **Patient CMV Status** | | | |
| CMV Negative | 48 (52.2%) | 21 (46.7%) | 69 (50.4%) |
| CMV Positive | 44 (47.8%) | 24 (53.3%) | 68 (49.6%) |
| **Recruitment Center** | | | |
| OSU | 48 (52.2%) | 28 (62.2%) | 76 (55.5%) |
| Alfred | 14 (15.2%) | 3 (6.7%) | 17 (12.4%) |
| St. Vincent | 18 (19.6%) | 5 (11.1%) | 23 (16.8%) |
| Hahnemann | 12 (13.0%) | 9 (20.0%) | 21 (15.3%) |
| **Prophylactic Methotrexate Use** | | | |
| No | 60 (65.2%) | 37 (82.2%) | 97 (70.8%) |
| Yes | 32 (34.8%) | 8 (17.8%) | 40 (29.2%) |

Table 5: Univariate Cox Regression Results for Disease-Free Survival

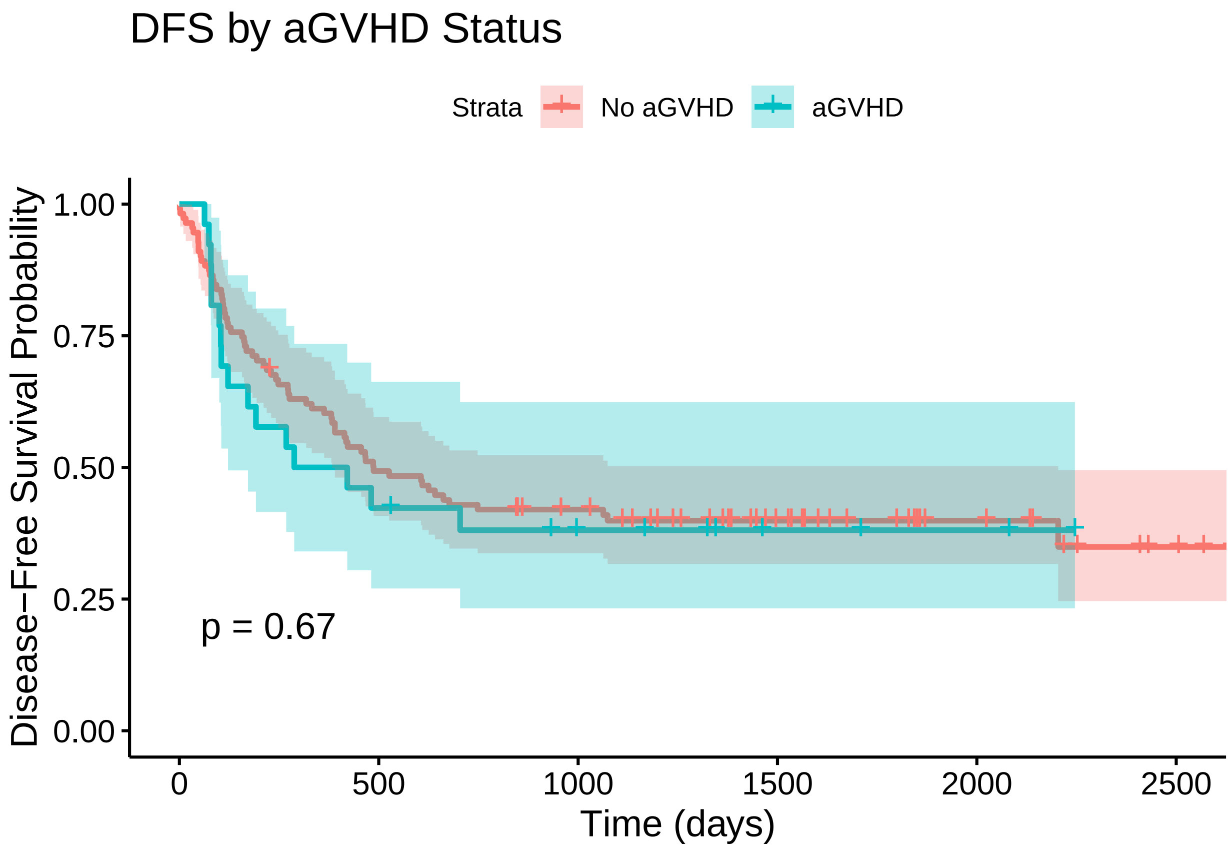
| variable | term | estimate | conf.low | conf.high | p.value |
|---|---|---:|---:|---:|---:|
| age | age | 1.011 | 0.988 | 1.035 | 0.338 |
| male | male | 0.795 | 0.514 | 1.228 | 0.301 |
| cmv | cmv | 1.167 | 0.759 | 1.796 | 0.482 |
| as.factor(disgroup) | as.factor(disgroup)2 | 0.563 | 0.321 | 0.989 | 0.046 |
| as.factor(disgroup) | as.factor(disgroup)3 | 1.467 | 0.869 | 2.478 | 0.152 |
| donorage | donorage | 1.014 | 0.990 | 1.040 | 0.252 |
| donormale | donormale | 0.991 | 0.633 | 1.552 | 0.970 |
| donorcmv | donorcmv | 1.047 | 0.677 | 1.620 | 0.836 |
| waittime | waittime | 1.000 | 0.999 | 1.001 | 0.791 |
| as.factor(hospital) | as.factor(hospital)2 | 2.122 | 1.148 | 3.922 | 0.016 |
| as.factor(hospital) | as.factor(hospital)3 | 0.912 | 0.495 | 1.681 | 0.768 |
| as.factor(hospital) | as.factor(hospital)4 | 0.421 | 0.191 | 0.930 | 0.032 |
| mtx | mtx | 1.489 | 0.934 | 2.373 | 0.094 |

### 4. It is generally thought that aGVHD has an anti-leukemic effect. Based on the available data, is occurrence of aGVHD after transplantation associated with improved disease-free survival? Is it associated with a decreased risk of relapse? In view of this, do you consider aGVHD as an important prognostic event?

```r
dfs_surv <- Surv(time = bmt$tdfs, event = bmt$deltadfs)

relapse_surv <- Surv(time = bmt$tdfs, event = bmt$deltar)

## univariate Cox for disease-free survival
cox_dfs_unadj <- coxph(dfs_surv ~ deltaa + age + as.factor(hospital) + male, data = bmt)
dfs_summary <- tidy(cox_dfs_unadj, conf.int = TRUE, exponentiate = TRUE)
dfs_summary
```

```
## # A tibble: 6 x 7
##   term                 estimate std.error statistic p.value conf.low conf.high
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 deltaa                  1.24     0.288      0.742  0.458     0.704      2.18
## 2 age                     1.02     0.0129     1.44   0.150     0.993      1.04
## 3 as.factor(hospital)2    1.88     0.328      1.93   0.0541    0.989      3.58
## 4 as.factor(hospital)3    0.805    0.325     -0.668  0.504     0.425      1.52
## 5 as.factor(hospital)4    0.342    0.423     -2.54   0.0112    0.149      0.783
## 6 male                    0.828    0.234     -0.807  0.419     0.524      1.31
```

```r
km_dfs <- survfit(dfs_surv ~ deltaa, data = bmt)

ggsurvplot(km_dfs, conf.int = TRUE, pval = TRUE,
           xlab = "Time (days)",
           ylab = "Disease-Free Survival Probability",
           legend.labs = c("No aGVHD", "aGVHD"),
           title = "DFS by aGVHD Status")
```



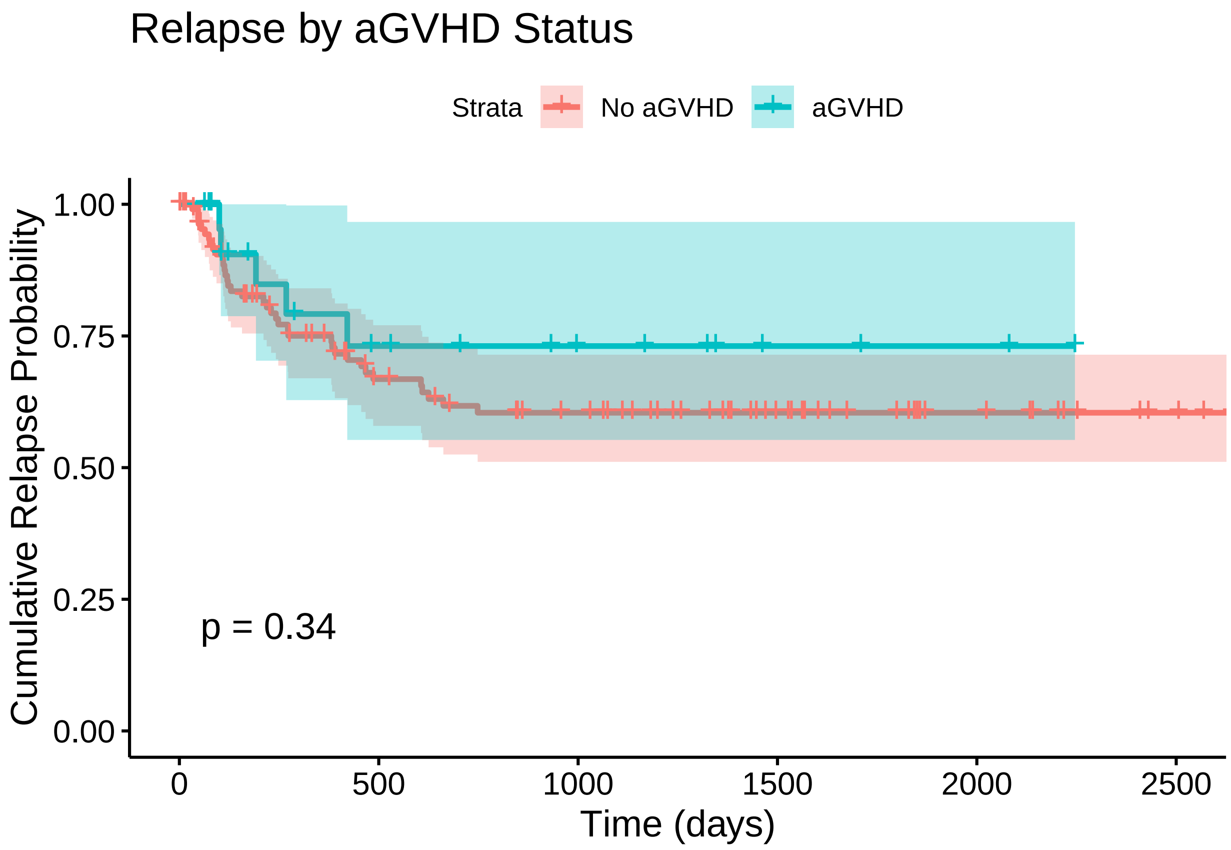
```
## multivariable Cox for dfs
cox_dfs_adj <- coxph(dfs_surv ~ deltaa + age + as.factor(hospital) + male, data = bmt)
dfs_summary_adj <- tidy(cox_dfs_adj, conf.int = T, exponentiate = T)
dfs_summary_adj
```

```
## # A tibble: 6 x 7
##   term                 estimate std.error statistic p.value conf.low conf.high
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 deltaa                  1.24     0.288      0.742  0.458     0.704     2.18
## 2 age                     1.02     0.0129     1.44   0.150     0.993     1.04
## 3 as.factor(hospital)2    1.88     0.328      1.93   0.0541    0.989     3.58
## 4 as.factor(hospital)3    0.805    0.325     -0.668  0.504     0.425     1.52
## 5 as.factor(hospital)4    0.342    0.423     -2.54   0.0112    0.149     0.783
## 6 male                    0.828    0.234     -0.807  0.419     0.524     1.31
```

```
## univariate Cox for relapse
cox_relapse_unadj <- coxph(relapse_surv ~ deltaa, data = bmt)
relapse_summary <- tidy(cox_relapse_unadj, conf.int = TRUE, exponentiate=TRUE)
relapse_summary
```

```
## # A tibble: 1 x 7
##   term   estimate std.error statistic p.value conf.low conf.high
##   <chr>     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 deltaa    0.638     0.477    -0.944   0.345    0.250      1.62
```

```
km_relapse <- survfit(relapse_surv ~ deltaa, data = bmt)

ggsurvplot(km_relapse, conf.int = TRUE, pval = TRUE,
           xlab = "Time (days)",
           ylab = "Cumulative Relapse Probability",
           legend.labs = c("No aGVHD", "aGVHD"),
           title = "Relapse by aGVHD Status")
```



```
## multivariable Cox for relapse
cox_relapse_adj <- coxph(relapse_surv ~ deltaa + age + as.factor(hospital) + male, data = bmt)
relapse_summary_adj <- tidy(cox_relapse_adj, conf.int = T, exponentiate = T)
relapse_summary_adj
```

```
## # A tibble: 6 x 7
##   term                  estimate std.error statistic p.value conf.low conf.high
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 deltaa                   0.682    0.491     -0.778   0.437    0.260      1.79
## 2 age                      1.01     0.0182     0.587   0.557    0.975      1.05
## 3 as.factor(hospital)2     1.21     0.555      0.350   0.726    0.409      3.60
## 4 as.factor(hospital)3     1.11     0.413      0.243   0.808    0.492      2.48
## 5 as.factor(hospital)4     0.588    0.513     -1.04    0.300    0.215      1.61
## 6 male                     0.673    0.320     -1.24    0.216    0.359      1.26
```

## 5. Among the patients who develop aGVHD, are any of the measured baseline factors associated with differences in disease-free survival?

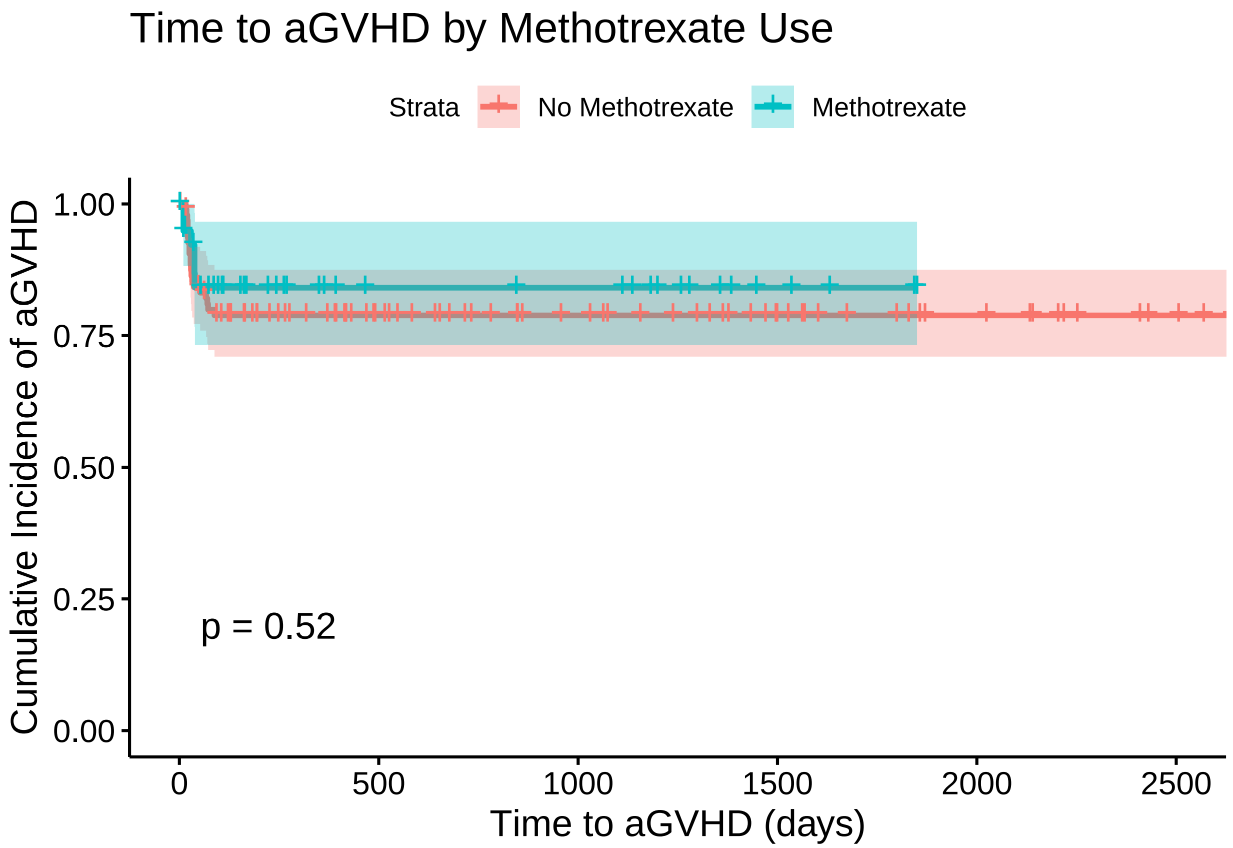
```r
bmt_agvhd <- subset(bmt, deltaa == 1)

dfs_surv_agvhd <- Surv(time = bmt_agvhd$tdfs, event = bmt_agvhd$deltadfs)

baseline_vars <- c("age", "male", "cmv", "as.factor(disgroup)",
                   "donorage", "donormale", "donorcmv",
                   "waittime", "as.factor(hospital)", "mtx")

uni_agvhd <- lapply(baseline_vars, function(var) {
    formula <- as.formula(paste("dfs_surv_agvhd ~", var))
    cox_model <- coxph(formula, data = bmt_agvhd)
    result <- tidy(cox_model, conf.int = TRUE, exponentiate=TRUE)
    result$variable <- var
    return(result)
})

uni_agvhd <- do.call(rbind, uni_agvhd)

uni_agvhd <- uni_agvhd[, c("variable", "term", "estimate", "conf.low", "conf.high", "p.value")]
kable(uni_agvhd, digits = 3, caption = "Univariate Cox Regression Results for DFS (aGVHD Patients Only)
```

Table 6: Univariate Cox Regression Results for DFS (aGVHD Patients Only)

| variable | term | estimate | conf.low | conf.high | p.value |
|---|---|---:|---:|---:|---:|
| age | age | 1.024 | 0.978 | 1.073 | 0.307 |
| male | male | 1.323 | 0.478 | 3.661 | 0.590 |
| cmv | cmv | 0.787 | 0.295 | 2.102 | 0.632 |
| as.factor(disgroup) | as.factor(disgroup)2 | 1.450 | 0.424 | 4.966 | 0.554 |
| as.factor(disgroup) | as.factor(disgroup)3 | 3.854 | 1.022 | 14.531 | 0.046 |
| donorage | donorage | 1.072 | 0.996 | 1.153 | 0.063 |
| donormale | donormale | 0.837 | 0.311 | 2.255 | 0.725 |
| donorcmv | donorcmv | 1.878 | 0.647 | 5.453 | 0.247 |
| waittime | waittime | 1.000 | 1.000 | 1.001 | 0.375 |
| as.factor(hospital) | as.factor(hospital)2 | 1.147 | 0.251 | 5.239 | 0.859 |
| as.factor(hospital) | as.factor(hospital)3 | 1.014 | 0.223 | 4.612 | 0.986 |
| as.factor(hospital) | as.factor(hospital)4 | 0.118 | 0.015 | 0.929 | 0.042 |
| mtx | mtx | 1.718 | 0.551 | 5.350 | 0.351 |

**6. Is prophylactic use of methotrexate associated with an increased or decreased risk of developing aGVHD? Provide an estimate of the survival function of time from transplant until onset of aGVHD separately for patients either administered methotrexate or not. In doing so, consider the importance of accounting for relevant confounding factors.**

```
agvhd_surv <- Surv(time = bmt$ta, event = bmt$deltaa)

## unadjusted KM
km_agvhd <- survfit(agvhd_surv ~ mtx, data = bmt)

ggsurvplot(km_agvhd, conf.int = TRUE, pval = TRUE,
           xlab = "Time to aGVHD (days)",
           ylab = "Cumulative Incidence of aGVHD",
           legend.labs = c("No Methotrexate", "Methotrexate"),
           title = "Time to aGVHD by Methotrexate Use")
```



```
## univariate Cox
cox_mtx <- coxph(agvhd_surv ~ mtx, data = bmt)
univariate_mtx <- tidy(cox_mtx, conf.int = TRUE, exponentiate=TRUE)
```

```
kable(univariate_mtx[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Univariate Cox Regression: Methotrexate and aGVHD Risk")
```

Table 7: Univariate Cox Regression: Methotrexate and aGVHD Risk

| term | estimate | conf.low | conf.high | p.value |
|---|---:|---:|---:|---:|
| mtx | 0.742 | 0.298 | 1.847 | 0.521 |

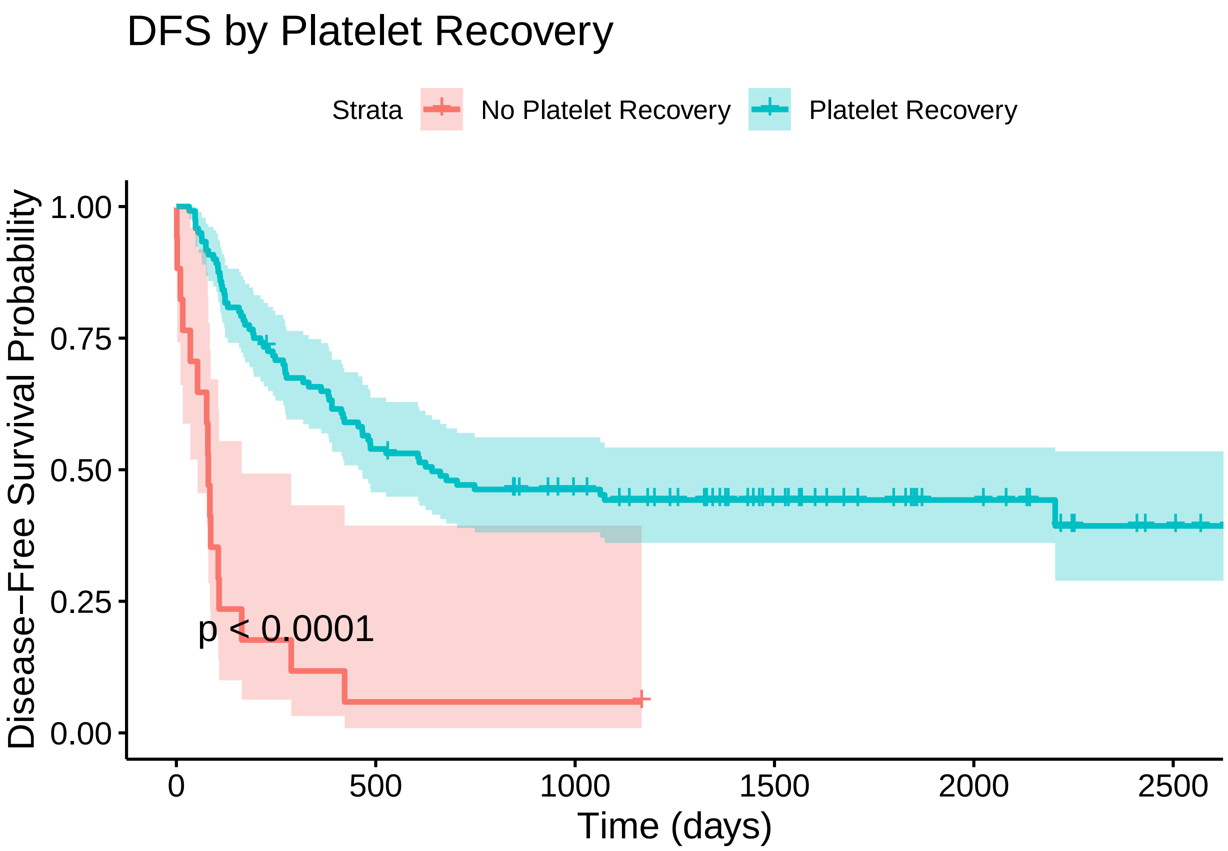
```
## multivariable Cox w/ confounders
cox_adj <- coxph(agvhd_surv ~ mtx + age + male + as.factor(disgroup) + waittime + as.factor(hospital), 
multivariable_mtx <- tidy(cox_adj, conf.int = TRUE, exponentiate=TRUE)

kable(multivariable_mtx[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Multivariable Cox Regression: Methotrexate and aGVHD Risk")
```

Table 8: Multivariable Cox Regression: Methotrexate and aGVHD Risk

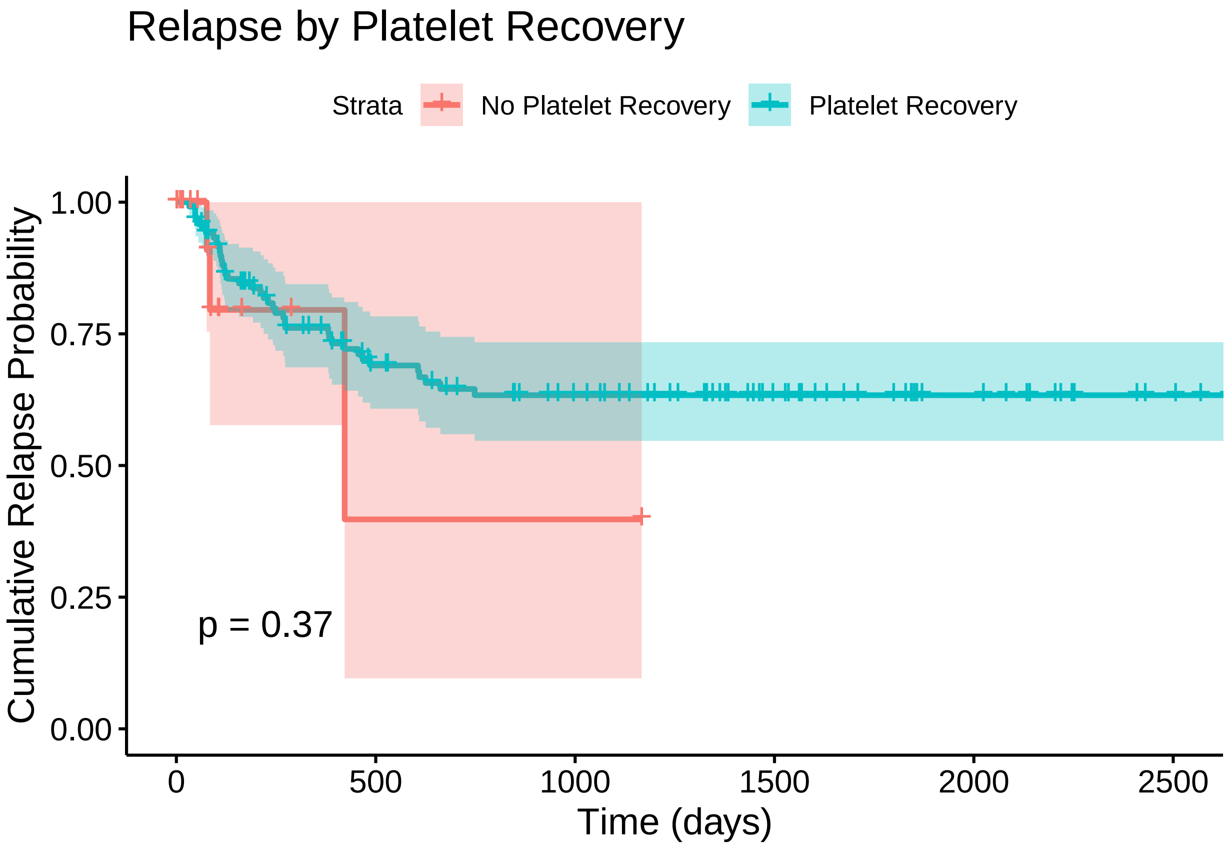
| term | estimate | conf.low | conf.high | p.value |
|---|---:|---:|---:|---:|
| mtx | 0.470 | 0.131 | 1.682 | 0.246 |
| age | 1.074 | 1.027 | 1.123 | 0.002 |
| male | 0.882 | 0.390 | 1.996 | 0.764 |
| as.factor(disgroup)2 | 0.528 | 0.177 | 1.580 | 0.254 |
| as.factor(disgroup)3 | 0.303 | 0.093 | 0.986 | 0.047 |
| waittime | 1.000 | 0.999 | 1.001 | 0.470 |
| as.factor(hospital)2 | 1.632 | 0.316 | 8.417 | 0.558 |
| as.factor(hospital)3 | NA | NA | NA | NA |
| as.factor(hospital)4 | 1.606 | 0.563 | 4.586 | 0.376 |

## 7. Based on the available data, is recovery of normal platelet levels associated with improved disease- free survival? Is it associated with a decreased risk of relapse?

```
dfs_surv <- Surv(time = bmt$tdfs, event = bmt$deltadfs)

relapse_surv <- Surv(time = bmt$tdfs, event = bmt$deltar)

## KM for dfs
km_dfs <- survfit(dfs_surv ~ deltap, data = bmt)

ggsurvplot(km_dfs, conf.int = TRUE, pval = TRUE,
           xlab = "Time (days)",
           ylab = "Disease-Free Survival Probability",
           legend.labs = c("No Platelet Recovery", "Platelet Recovery"),
           title = "DFS by Platelet Recovery")
```



```r
## KM for relapse
km_relapse <- survfit(relapse_surv ~ deltap, data = bmt)

ggsurvplot(km_relapse, conf.int = TRUE, pval = TRUE,
           xlab = "Time (days)",
           ylab = "Cumulative Relapse Probability",
           legend.labs = c("No Platelet Recovery", "Platelet Recovery"),
           title = "Relapse by Platelet Recovery")
```



```r
## univariate Cox for dfs
cox_dfs <- coxph(dfs_surv ~ deltap, data = bmt)
univariate_dfs <- tidy(cox_dfs, conf.int = TRUE, exponentiate=TRUE)

kable(univariate_dfs[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Univariate Cox: Platelet Recovery and DFS")
```

Table 9: Univariate Cox: Platelet Recovery and DFS

| term | estimate | conf.low | conf.high | p.value |
|---|---|---|---|---|
| deltap | 0.188 | 0.107 | 0.33 | 0 |

```r
## multivariable Cox for dfs
cox_dfs_adj <- coxph(dfs_surv ~ deltap + age + as.factor(hospital) + as.factor(disgroup) + male, data =
multi_dfs <- tidy(cox_dfs_adj, conf.int = T, exponentiate = T)

kable(multi_dfs[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Multivariable Cox: Platelet Recovery and DFS")
```

Table 10: Multivariable Cox: Platelet Recovery and DFS

| term | estimate | conf.low | conf.high | p.value |
|:---|---:|---:|---:|---:|
| deltap | 0.256 | 0.138 | 0.476 | 0.000 |
| age | 1.017 | 0.992 | 1.044 | 0.189 |
| as.factor(hospital)2 | 1.347 | 0.662 | 2.739 | 0.411 |
| as.factor(hospital)3 | 0.869 | 0.453 | 1.667 | 0.672 |
| as.factor(hospital)4 | 0.477 | 0.206 | 1.102 | 0.083 |
| as.factor(disgroup)2 | 0.702 | 0.375 | 1.312 | 0.267 |
| as.factor(disgroup)3 | 1.496 | 0.835 | 2.680 | 0.176 |
| male | 0.928 | 0.580 | 1.484 | 0.756 |

```r
## univariate Cox for relapse
cox_relapse <- coxph(relapse_surv ~ deltap, data = bmt)
univariate_relapse <- tidy(cox_relapse, conf.int = TRUE)

kable(univariate_relapse[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Univariate Cox: Platelet Recovery and Relapse Risk")
```

Table 11: Univariate Cox: Platelet Recovery and Relapse Risk

| term | estimate | conf.low | conf.high | p.value |
|:---|---:|---:|---:|---:|
| deltap | -0.539 | -1.729 | 0.65 | 0.374 |

```r
## multivariable Cox for relapse
cox_relapse_adj <- coxph(relapse_surv ~ deltap + age + as.factor(hospital) + + as.factor(disgroup)+ male
multi_relapse <- tidy(cox_relapse_adj, conf.int = T, exponentiate = T)

kable(multi_relapse[, c("term", "estimate", "conf.low", "conf.high", "p.value")],
      digits = 3, caption = "Multivariable Cox: Platelet Recovery and DFS")
```

Table 12: Multivariable Cox: Platelet Recovery and DFS

| term | estimate | conf.low | conf.high | p.value |
|:---|---:|---:|---:|---:|
| deltap | 0.895 | 0.256 | 3.134 | 0.863 |
| age | 0.998 | 0.960 | 1.038 | 0.937 |
| as.factor(hospital)2 | 1.100 | 0.359 | 3.374 | 0.868 |
| as.factor(hospital)3 | 1.324 | 0.575 | 3.051 | 0.509 |
| as.factor(hospital)4 | 0.783 | 0.274 | 2.241 | 0.649 |
| as.factor(disgroup)2 | 0.415 | 0.163 | 1.057 | 0.065 |
| as.factor(disgroup)3 | 1.836 | 0.821 | 4.106 | 0.139 |
| male | 0.637 | 0.335 | 1.213 | 0.170 |