Biostat 215 Project

Cindy Pang, Jon Hori, Sophie Phillips

2025-03-02

```
library(survival)
library(dplyr)
```

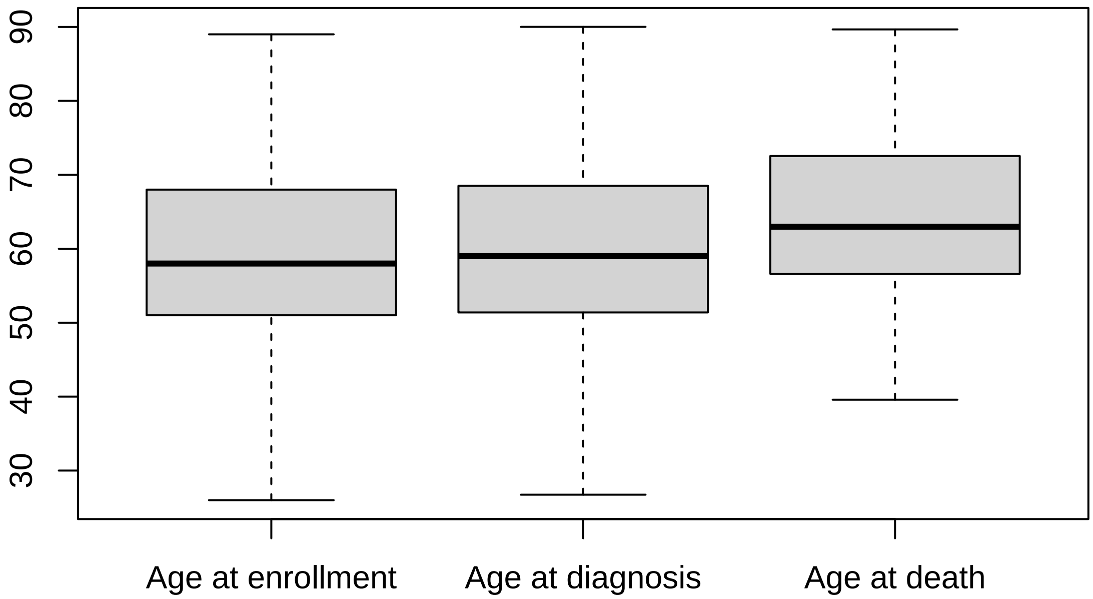
```
##
## Attaching package: 'dplyr'

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

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

```
dat <- (read.csv("../data/survProjData.csv")
        %>% dplyr::select(sample, id, race = race.demographic,
                          ethicity = ethnicity.demographic,
                          vital_status = vital_status.demographic,
                          age_at_index = age_at_index.demographic,
                          days_to_birth = days_to_birth.demographic,
                          year_of_birth = year_of_birth.demographic,
                          days_to_death = days_to_death.demographic,
                          year_of_death = year_of_death.demographic,
                          figo_stage.diagnoses, days_to_last_follow_up.diagnoses,
                          age_at_diagnosis = age_at_diagnosis.diagnoses,
                          age_at_diagnosis_years = age_at_earliest_diagnosis_in_years.diagnoses.xena_der
                          primary_diagnosis.diagnoses,
                          shortest_dimension.samples, intermediate_dimension.samples,
                          longest_dimension.samples,
                          sample_type.samples))
dat$age_at_index_days <- dat$age_at_index * 365
dim(dat)
```

```
## [1] 609  20
```

We have data on 609 female patients. Ovarian tissue samples were collected regularly and analyzed for cancer. All subjects in the data set were diagnosed with ovarian cancer, with the majority (600) having Serous cystadenocarcinoma. The data are left-truncated by age of enrollment. 75% of patients enrolled after age 50.

```
boxplot(dat$age_at_index, dat$age_at_diagnosis_years, dat$age_at_index + (dat$days_to_death/365),
        names = c("Age at enrollment", "Age at diagnosis", "Age at death"))
```



Diagnoses occurred between ages 26-90. Four subjects were missing both time-to-death data and time-to-diagnosis data. We removed these patients. Seven subjects were missing time-to diagnosis data but had time-to-death data. We included these subjects in the time-to-death analysis.

```
dat <- dat %>% filter(!is.na(age_at_diagnosis) | !is.na(days_to_death))
dim(dat)
```

```
## [1] 605  20
```

366 of the subjects died during the study.

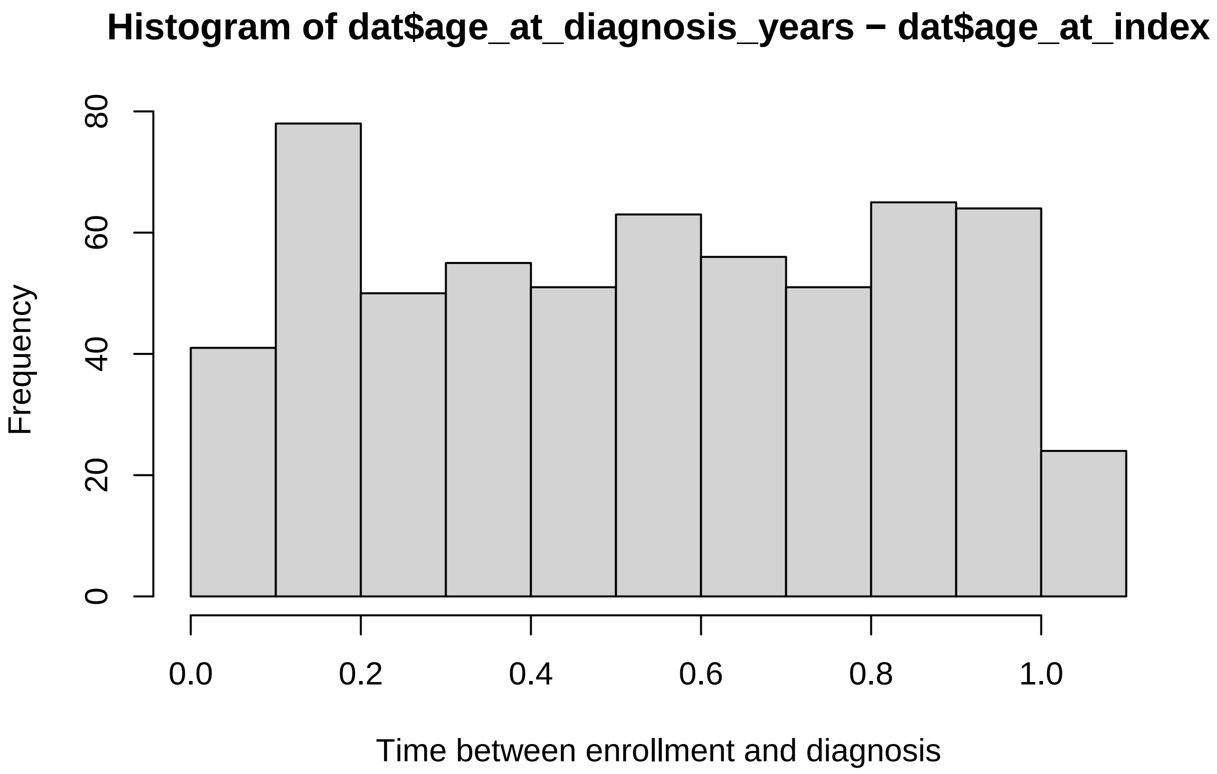
```
summary(dat$days_to_death)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       8     592    1078    1171    1579    4624     240
```

We will analyze both the time to diagnosis and time to death with a competing hazards method. The time to diagnosis is uncensored as all subjects were diagnosed within the study but the time to death is censored as only half the subjects died.

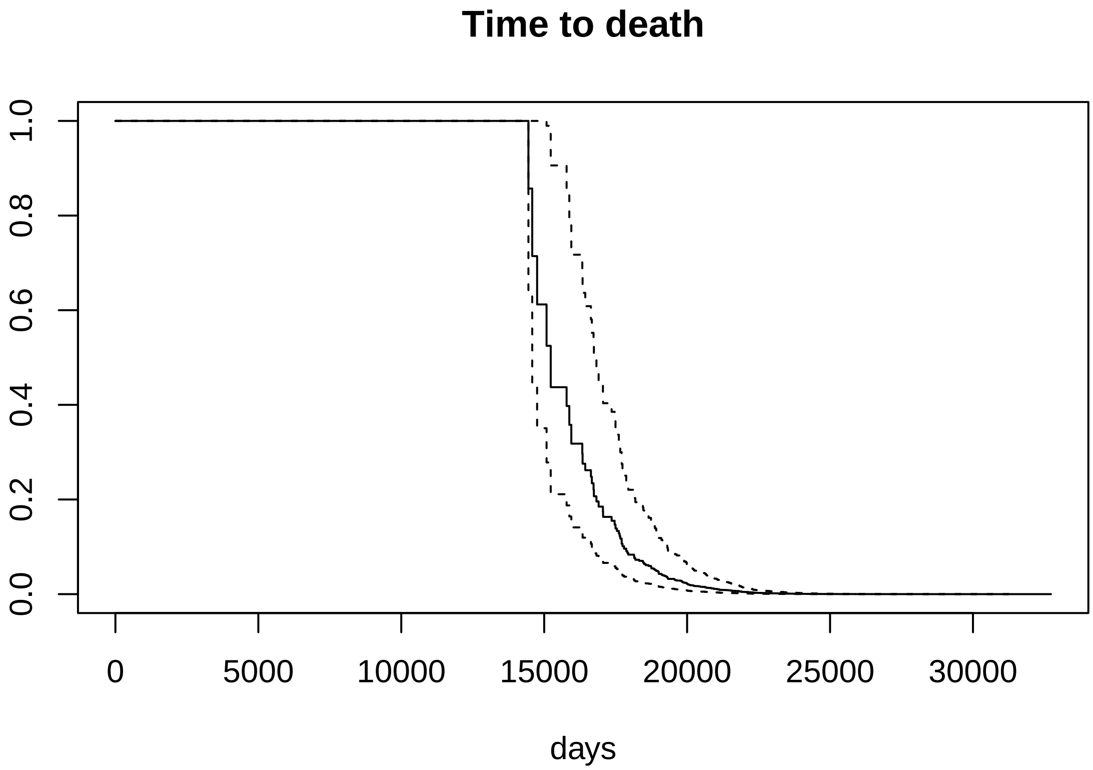
It is not clear how we should handle the left truncation. The time between enrollment and diagnosis is small causing the Kaplan-Meier curve to fall to 0 immediately.

```
hist(dat$age_at_diagnosis_years - dat$age_at_index, xlab = "Time between enrollment and diagnosis")
```



```r
# fit <- survfit(Surv(dat$age_at_index, dat$age_at_diagnosis_years, rep(1, nrow(dat))) ~ 1,
#                      data = dat)
# plot(fit, main = "Time to diagnosis")
```

```r
fit <- survfit(Surv(age_at_index_days, age_at_index_days+days_to_death, obs) ~ 1,
                    data = dat %>% mutate(obs = vital_status == "Dead"))
plot(fit, main = "Time to death", xlab = "days")
```



We will address the following questions:

1. Does the time of cancer onset vary between different races? The majority (521) of subjects were white, so we may have limited power to detect differences. We will approach this question with a log-rank test to compare survival curves.

```
table(dat$race)
```

```
##
##          american indian or alaska native
##                                         3
##                                     asian
##                                        22
##                 black or african american
##                                        34
## native hawaiian or other pacific islander
##                                         1
##                              not reported
##                                        26
##                                     white
##                                       519
```

2. Does the diagnosed FIGO stage predict the time to death? In theory, higher stage cancers should have shorter times to death. We will analyze this with a multi-group log rank test.

3. We will use a Cox proportional hazards model to determine how the size of the collected tissue sample relates to the risk of death.

```r
summary(dat$intermediate_dimension.samples)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##   0.300   0.600   0.800   0.885   1.000   3.000      20
```