# S3T4

Group A3

2025-11-05

### Step 1: Joining datasets

```
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
```

```
library(ggplot2)
library(readr)
library(tableone)
```

```
## Warning: package 'tableone' was built under R version 4.5.2
```

```
library(broom)
library(ResourceSelection) # Hosmer-Lemeshow test
```

```
## Warning: package 'ResourceSelection' was built under R version 4.5.2
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.5.2
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(car) # VIF
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
library(survival) # time-to-event
library(cmprsk)
```

```
## Warning: package 'cmprsk' was built under R version 4.5.2
```

```
library(tidyr)
library(GGally)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(knitr)

data_t3 <- read_csv("Data_T3.csv")
```

```
## New names:
## * `` -> `...1`
```

```
## Rows: 300 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): TreatmentGroup, Sex
## dbl (7): ...1, ID, Age, ECOG_PS, GFR, Time, Event
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_t4 <- read_csv("Data_T4.csv")
```

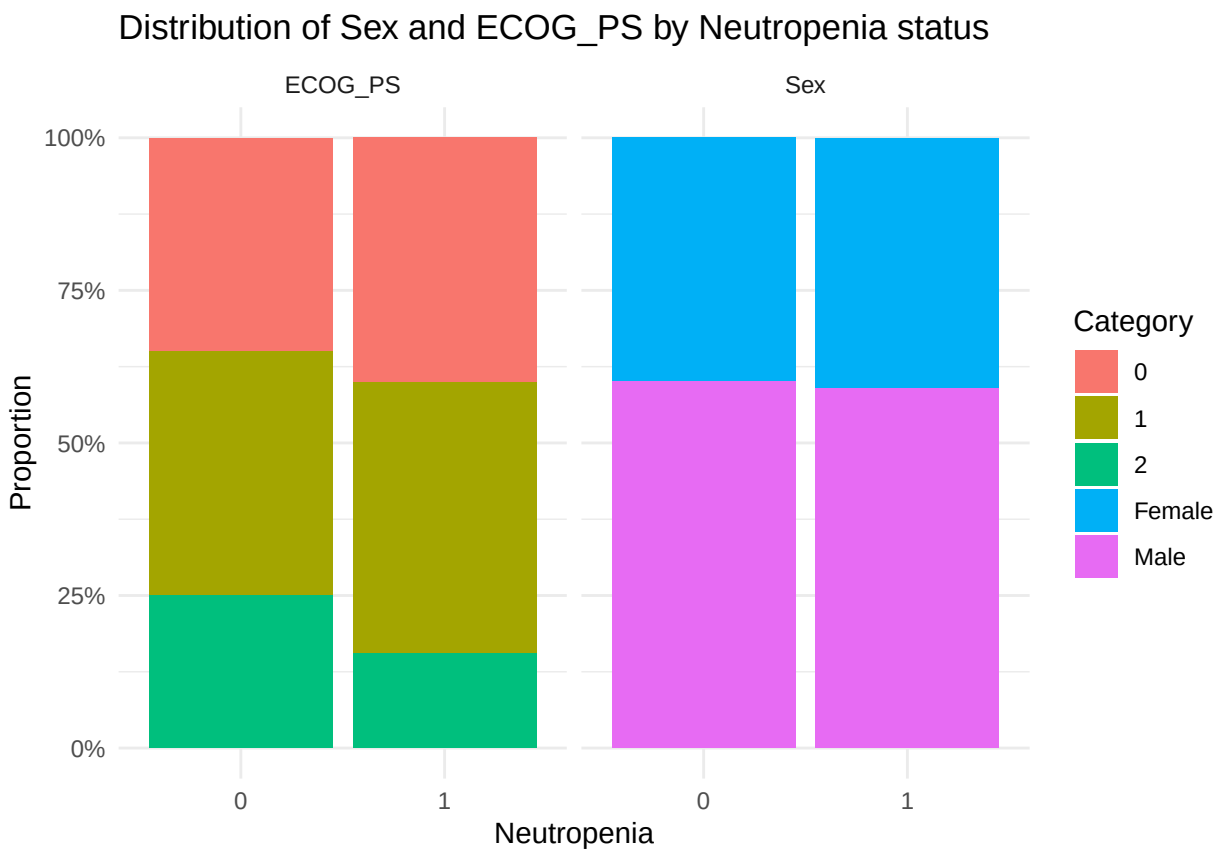
```
## New names:
## Rows: 150 Columns: 3
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (3): ...1, ID, Neutropenia
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

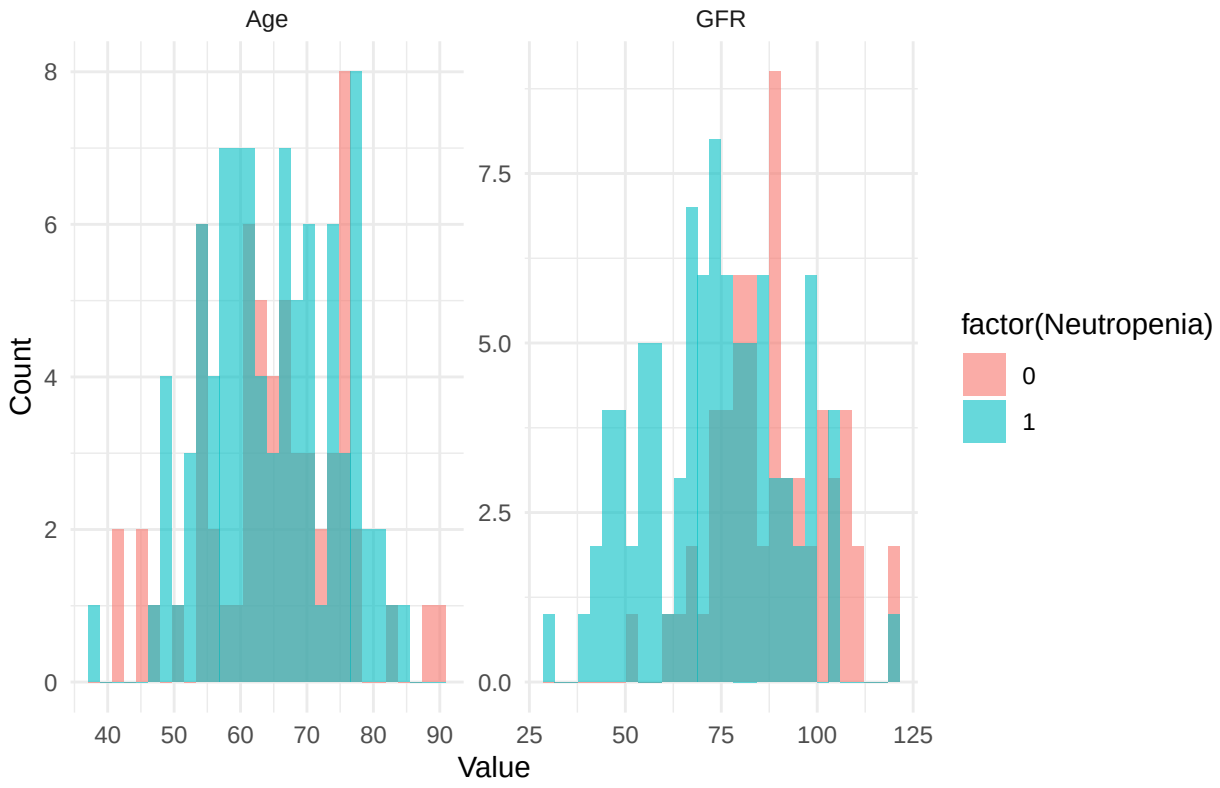
```r
treat_group <- data_t3 %>% filter(TreatmentGroup == "Treatment")
merged <- treat_group %>%
  left_join(data_t4, by = "ID")
```

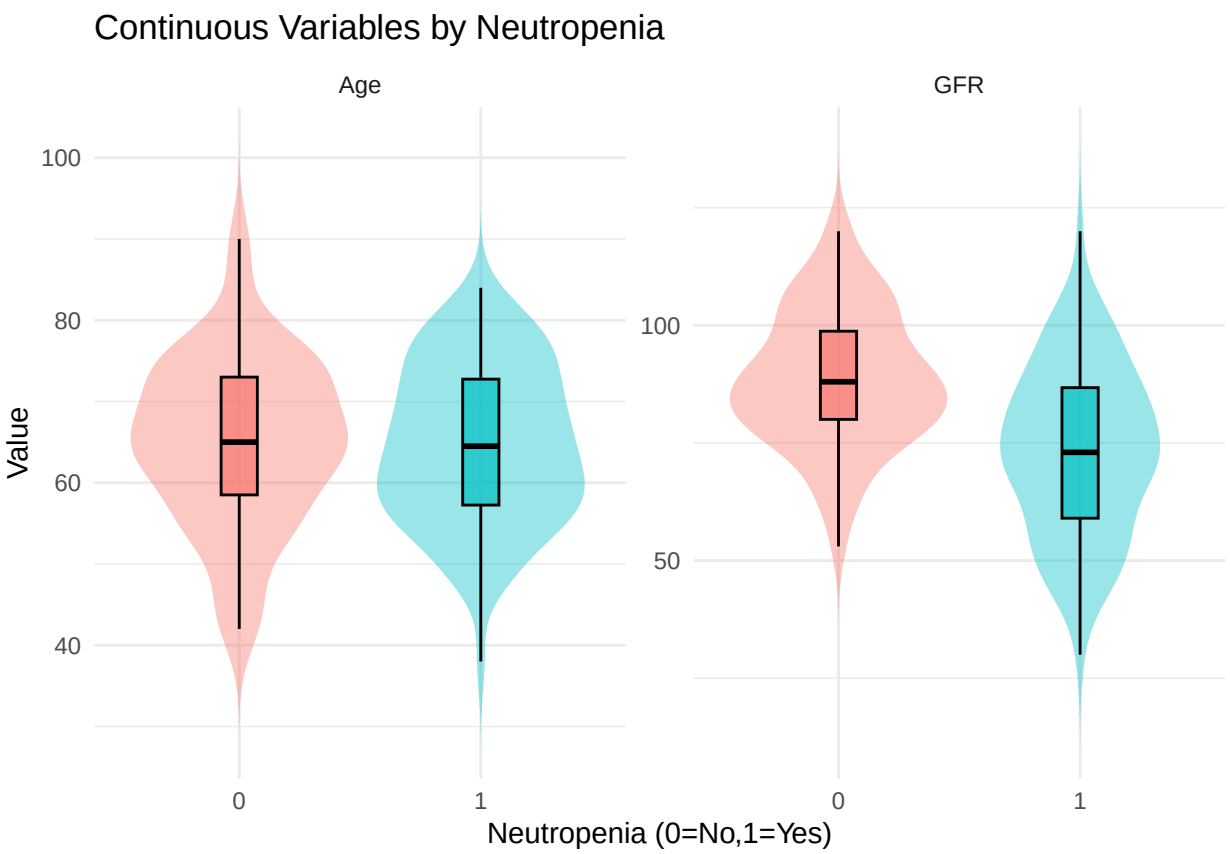
## Step 2: Data exploration

We examined the data, visualized the distribution of categorical data, including Sex and EcoG_PS, and continuous variables, including GFR and Age, by Neutropenia status, and the distribution of Neutropenia Status in categorical data.

```
##                   Stratified by Neutropenia
##                    level  0             1             p      test
##   n                            60            90
##   Age (mean (SD))         65.05 (10.46) 64.68 (9.71)   0.824
##   Sex (%)          Female     24 (40.0)     37 (41.1)   1.000
##                    Male       36 (60.0)     53 (58.9)
##   ECOG_PS (%)      0          21 (35.0)     36 (40.0)   0.356
##                    1          24 (40.0)     40 (44.4)
##                    2          15 (25.0)     14 (15.6)
##   GFR (mean (SD))         88.52 (14.50) 73.28 (18.09) <0.001
```

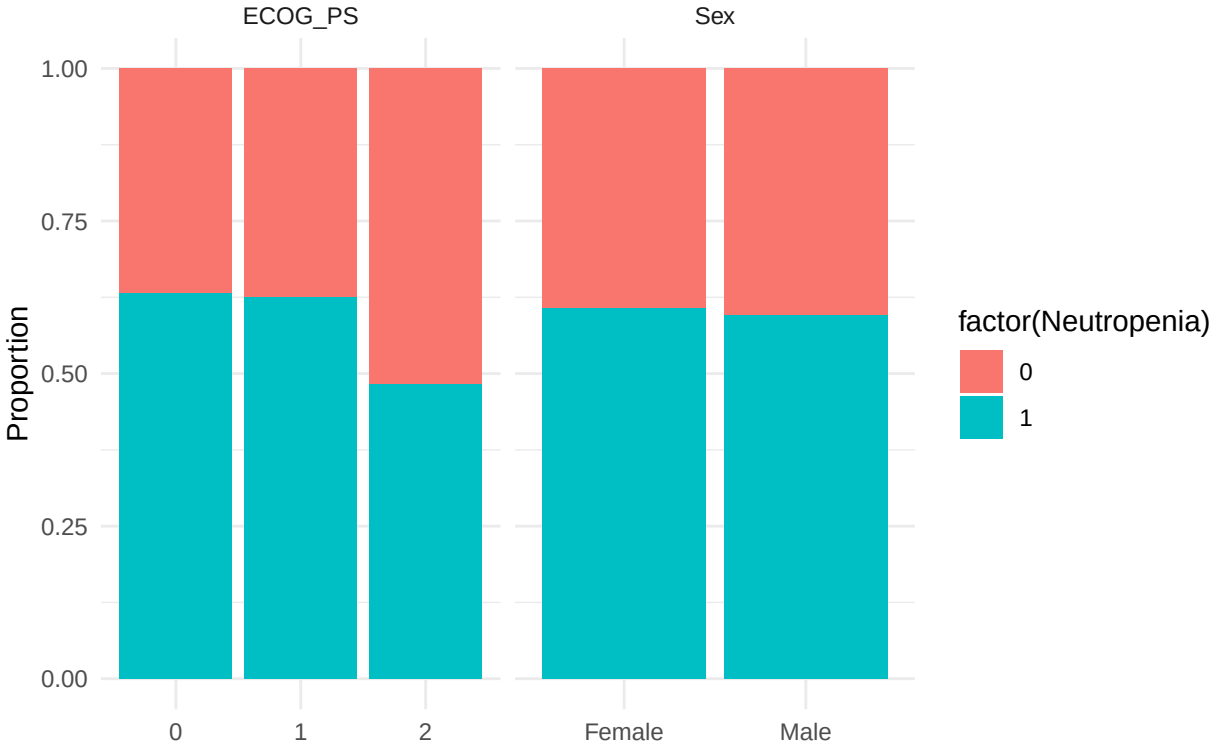

Distribution of Sex and ECOG_PS by Neutropenia status
ECOG_PS
Sex
100%
75%
50%
25%
0%
Proportion
0
1
0
1
Neutropenia
Category
0
1
2
Female
Male


Distribution of Continuous Variables by Neutropenia
Age
GFR
8
6
4
2
0
Count
7.5
5.0
2.5
0.0
40
50
60
70
80
90
25
50
75
100
125
Value
factor(Neutropenia)
0
1




```
## Warning: There were 2 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘text = text_fn(.data$x, .data$y)‘.
## i In group 1: ‘color = 0‘.
## Caused by warning in ‘cor()‘:
## ! the standard deviation is zero
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
## There were 2 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘text = text_fn(.data$x, .data$y)‘.
## i In group 1: ‘color = 0‘.
## Caused by warning in ‘cor()‘:
## ! the standard deviation is zero
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 1 remaining warning.
```


Categorical Variables by Neutropenia




Results shows that continuous variables are seemingly in a normal distribution, GRF distribution looks differently in different Neutropenia case, while Age doesn't seem to have significant influence on Neutropenia.

Percentage of high-level EcoG_PS is higher in Neutropenia cases, while sex distribute similarly in with and without Neutropenia.

Next, we carry out statistic tests on continuous variables.

```
##
##     Female Male
##   0     24   36
##   1     37   53
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(merged$Neutropenia, merged$Sex)
## X-squared = 0, df = 1, p-value = 1
```

```
##
##      0  1  2
##   0 21 24 15
##   1 36 40 14
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(merged$Neutropenia, merged$ECOG_PS)
## X-squared = 2.0644, df = 2, p-value = 0.3562
```

Table 1: Comparison of Continuous Variables by Neutropenia

| | Variable | Normality_Group0_p | Normality_Group1_p | Test | Statistic | p_value |
|---|---|---|---|---|---|---|
| t...1 | Age | 0.4044 | 0.2335 | t-test | 0.2196 | 0.8265 |
| t...2 | GFR | 0.8557 | 0.8876 | t-test | 5.7022 | 0.0000 |

From the results of the Pearson's Chi-squared tests, p-value of sex is 1, showing Neutropenia is statistically independent from Sex. The p-value for test of EcoG is 0.3562, higher than 0.05, which means that Neutropenia is also approximately independent from EcoG.

From Shapiro Tests, both continuous variables fit the normal distribution, which also accord with the histogram. While the p-value of age is higher than 0.05 showing significant difference in age with or without Neutropenia, the p-value of GFR is close to 0, showing no significant difference in GFR between different Neutropenia state. This result does not fit the violin plot and boxplot. The reason might be that the medium is shifting from mean value.

## Step 3: Modeling

```
##
## Call:
## glm(formula = Neutropenia ~ Age + Sex + ECOG_PS + GFR, family = binomial,
##     data = merged)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.73875    1.51573   3.126  0.00177 **
## Age          0.01113    0.01921   0.579  0.56242
## SexMale     -0.18031    0.38281  -0.471  0.63764
## ECOG_PS     -0.33155    0.25237  -1.314  0.18894
## GFR         -0.05752    0.01235  -4.658 3.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 201.90  on 149  degrees of freedom
## Residual deviance: 172.05  on 145  degrees of freedom
## AIC: 182.05
##
## Number of Fisher Scoring iterations: 4
```

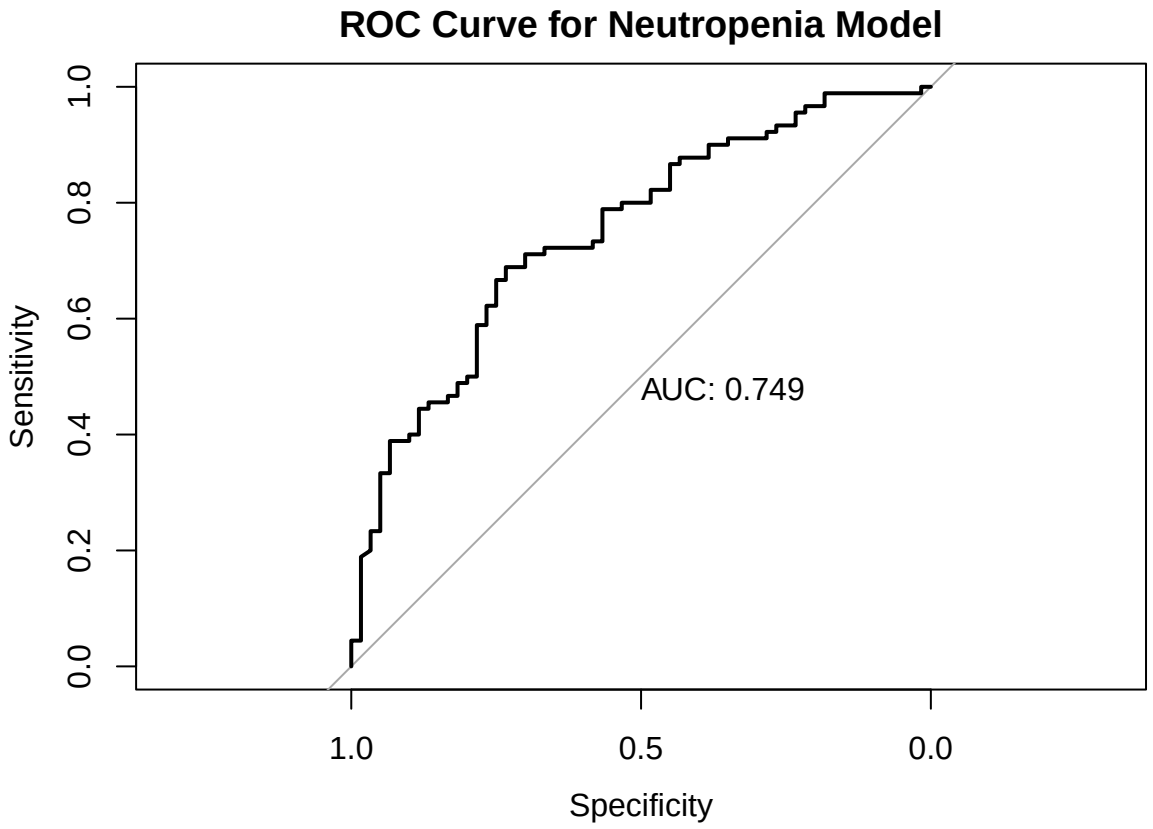
```
## # A tibble: 5 x 7
##   term        estimate std.error statistic    p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
## 1 (Intercept)  114.       1.52       3.13  0.00177       6.62   2608.
## 2 Age            1.01     0.0192     0.579 0.562         0.974     1.05
## 3 SexMale        0.835    0.383     -0.471 0.638         0.390     1.76
## 4 ECOG_PS        0.718    0.252     -1.31  0.189         0.434     1.17
## 5 GFR            0.944    0.0123    -4.66  0.00000319    0.920     0.966
```

```
##      Age      Sex  ECOG_PS      GFR
## 1.044859 1.028674 1.008695 1.040921
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  merged$Neutropenia, fitted(model_logit)
## X-squared = 8.6298, df = 8, p-value = 0.3745
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




We used logistic regression to establish a model for this case. The log-odds and odds-ratio shows that only GFR has a significant influence on Neutropenia state, better GFR leads to lower risk of Neutropenia. VIF value are all close to 1, which means that there are no obvious colinearality problem in this model. To access the performance, we accessed the AUC value, which is 0.749, showing the model is acceptable.

For treatment, it is suggested by this model that renal function can support treatment to a better result.

### Step 4: Future Work

In the future, maybe we can include factor of time of Neutropenia, to build a time-to-event model.