AKKODIS INTERVIEW

Analysis of an oncological dataset

Erik De Luca

SUMMARY

- Assignment description
- Setting the enviroment
- Descriptive Analysis
- Survival Analysis
- Deeper Analysis

DATASET DESCRIPTION

The response variable, `SurvTime`, is the survival time in days of a lung cancer patient.

The covariates are:

- `Cell` (type of cancer cell),
- `Therapy` (type of therapy: standard or test),
- `Prior` (prior therapy: 0=no, 10=yes),
- `Age` (age in years),
- `DiagTime`(time in months from diagnosis to entry into the trial)
- `Kps`(performance status).

A censoring indicator variable Censor is created from the data, with the value 1 indicating a censored time and the value 0 indicating an event time. Since there are only two types of therapy, an indicator variable, Treatment, is constructed for therapy type, with: value 0 for standard therapy and value 1 for test therapy.

EXERCISES

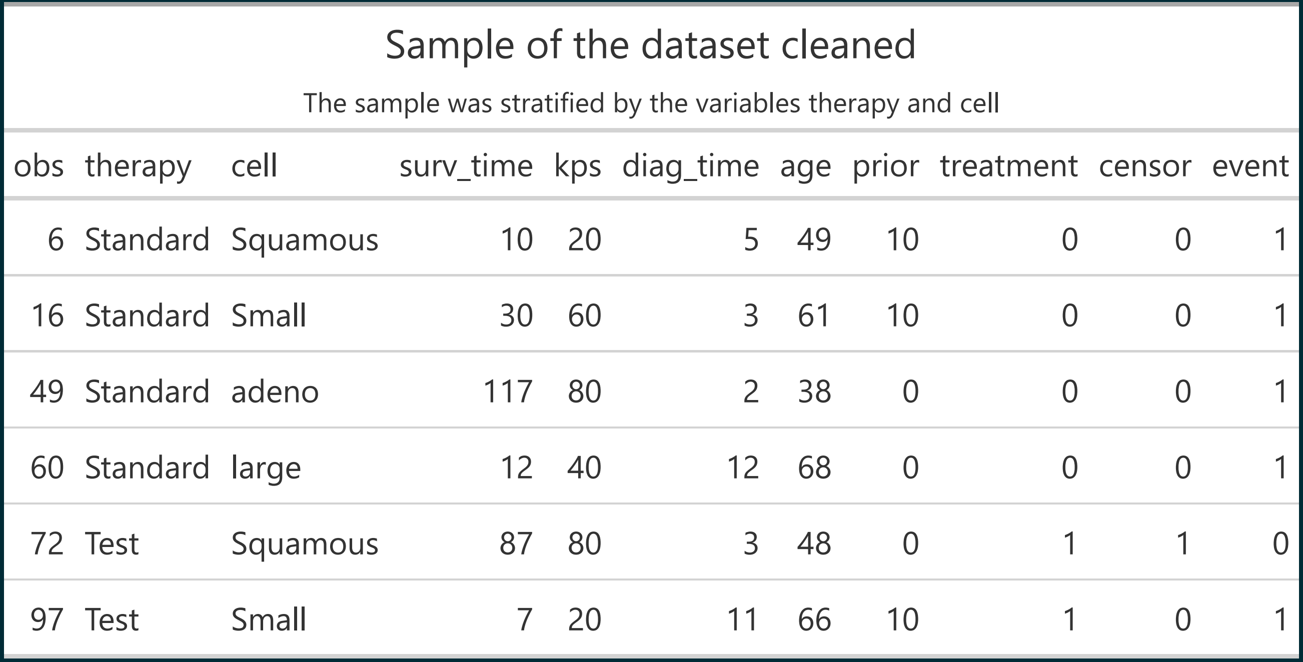
1. what was the maximum survival time for the cell type adeno?
2. what is the average age of subjects in this study?
3. which cell type appeared the most during this study?
4. Calculate descriptive statistics for all numeric variables within this dataset?
5. Perform a survival analysis to assess the survival time (variable SurvTime)? based on the cancerous cells (var Cell)? Consider applying survival functions/kaplan meier quartiles/cumulative incidence function/cox regression etc.
6. Perform an appropriate multivariable analysis to analyze the effect of independent variables age on the hazard ratio between the different cancerous cells (var Cell)?

LOAD LIBRARIES

```r
pacman::p_load(
  tidyverse,  # A set of many useful libraries
  readxl,     # To import the dataset from Excel
  here,       # To avoid problems with file directories
  janitor,    # To clean data in a fast way
  gt,         # Output tables
  gtsummary,  # Output tables for models and survival data
  survival,   # To manage survival data
  ggsurvfit,  # To plot survival analysis
  tidycmprsk  # To fit survival models
)
```

IMPORT DATA

```r
data_imported <- read_xlsx(
  here("Akkodis interview", "Oncology_dataset_for_R.xlsx")
  )

data_cleaned <- data_imported |>
  clean_names() |>
  remove_empty() |>
  remove_constant()

data_cleaned |>
  slice_sample(n = 1, by = c(therapy, cell)) |>
  gt() |>
  tab_header("Sample of the dataset cleaned",
             "The sample was stratified by the variables therapy and cell")
```



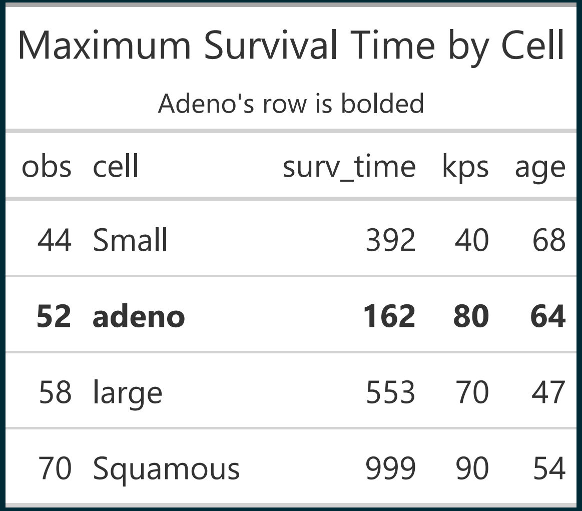
Sample of the dataset cleaned

The sample was stratified by the variables therapy and cell

obs	therapy	cell	surv_time	kps	diag_time	age	prior	treatment	censor	event
6	Standard	Squamous	10	20	5	49	10	0	0	1
16	Standard	Small	30	60	3	61	10	0	0	1
49	Standard	adeno	117	80	2	38	0	0	0	1
60	Standard	large	12	40	12	68	0	0	0	1
72	Test	Squamous	87	80	3	48	0	1	1	0
97	Test	Small	7	20	11	66	10	1	0	1

WHAT WAS THE MAXIMUM SURVIVAL TIME FOR THE CELL TYPE ADENO?

```r
data_cleaned |>
  filter(
    surv_time == max(surv_time),
    .by = cell
    ) |>
  select(obs, cell, surv_time, kps, age) |>
  gt() |>
  tab_header("Maximum Survival Time by Cell",
             "Adeno's row is bolded") |>
  tab_style(
    style = cell_text(weight = "bold"),
    locations = list(
      cells_body(rows = cell == "adeno")
      )
  )
```



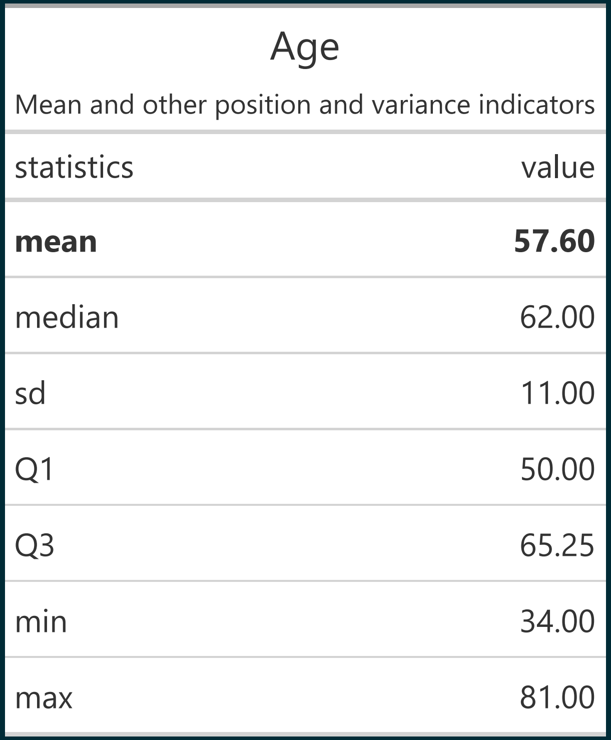
Maximum Survival Time by Cell

Adeno's row is bolded

obs	cell	surv_time	kps	age
44	Small	392	40	68
52	**adeno**	**162**	**80**	**64**
58	large	553	70	47
70	Squamous	999	90	54

WHAT IS THE AVERAGE AGE OF SUBJECTS IN THIS STUDY?

```r
data_cleaned |>
  summarise(
    across(age, list(
      mean = mean,
      median = median,
      sd = ~ round(sd(.)),
      Q1 = ~ quantile(., .25),
      Q3 = ~ quantile(., .75),
      min = min,
      max = max
    ),
    .names = "{fn}"
    )
  ) |>
  pivot_longer(everything(), names_to = "statistics") |>
  gt() |>
  tab_header(
    "Age",
    "Mean and other position and variance indicators"
    ) |>
  tab_style(
    style = cell_text(weight = "bold"),
    locations = list(
      cells_body(rows = statistics == "mean")
      )
```



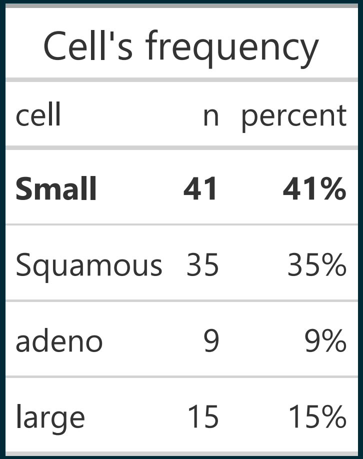
Age

Mean and other position and variance indicators

statistics	value
mean	**57.60**
median	62.00
sd	11.00
Q1	50.00
Q3	65.25
min	34.00
max	81.00

WHICH CELL TYPE APPEARED THE MOST DURING THIS STUDY?

```r
data_cleaned |>
  tabyl(cell) |>
  adorn_pct_formatting(digits = 0) |>
  gt() |>
  tab_header("Cell's frequency") |>
  tab_style(
    style = cell_text(weight = "bold"),
    locations = list(
      cells_body(rows = n == max(n))
      )
  )
```



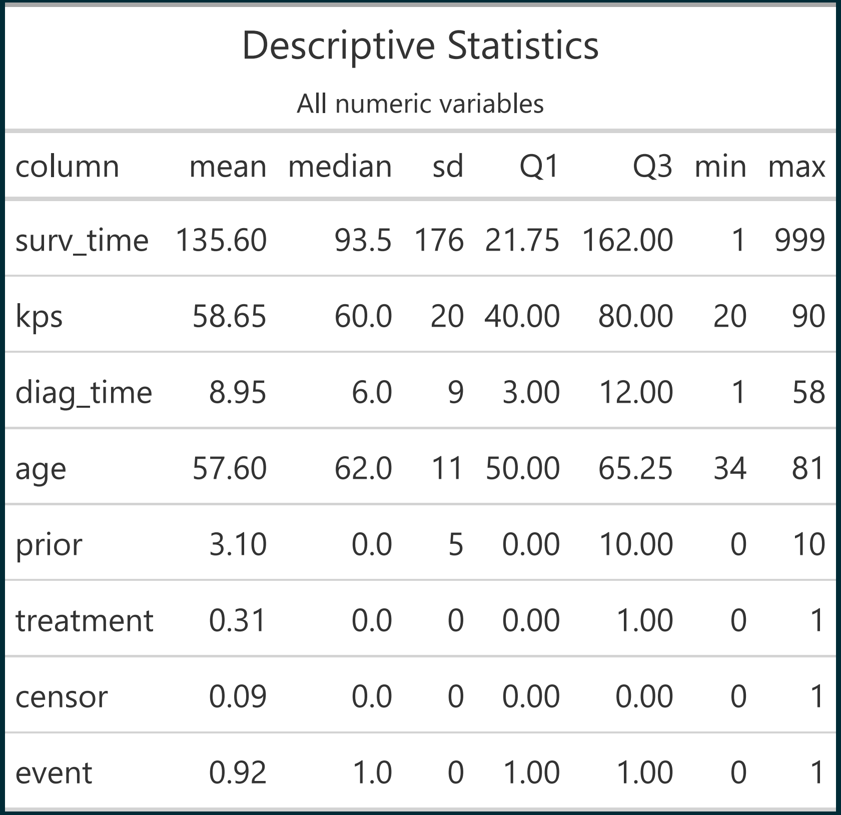
Cell's frequency

cell	n	percent
Small	**41**	**41%**
Squamous	35	35%
adeno	9	9%
large	15	15%

CALCULATE DESCRIPTIVE STATISTICS FOR ALL NUMERIC VARIABLES WITHIN THIS DATASET?

```r
data_cleaned |>
  select(-obs) |>
  summarise(
    across(
      where(is.numeric),
           list(
      mean = mean,
      median = median,
      sd = ~ round(sd(.)),
      Q1 = ~ quantile(., .25),
      Q3 = ~ quantile(., .75),
      min = min,
      max = max
    ),
    .names = "{col}-{fn}"
    )
  ) |>
  pivot_longer(everything(), names_to = "statistics") |>
  separate(col = statistics, sep = "-", into = c("column", "statistics")) |>
  pivot_wider(names_from = statistics, values_from = value) |>
  gt() |>
  tab_header("Descriptive Statistics", "All numeric variables")
```

CALCULATE DESCRIPTIVE STATISTICS FOR ALL NUMERIC VARIABLES WITHIN THIS DATASET?



Descriptive Statistics

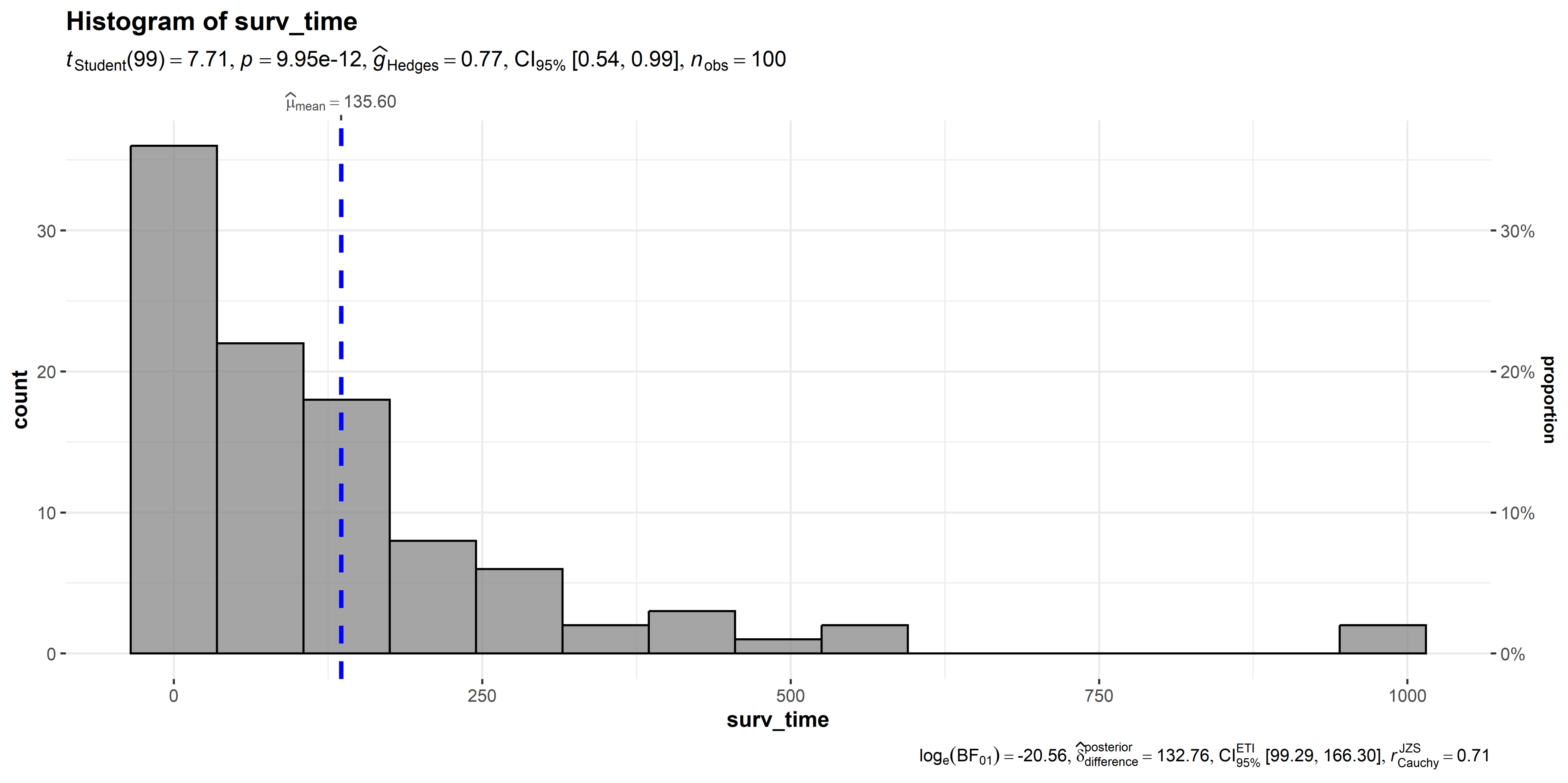
All numeric variables

column	mean	median	sd	Q1	Q3	min	max
surv_time	135.60	93.5	176	21.75	162.00	1	999
kps	58.65	60.0	20	40.00	80.00	20	90
diag_time	8.95	6.0	9	3.00	12.00	1	58
age	57.60	62.0	11	50.00	65.25	34	81
prior	3.10	0.0	5	0.00	10.00	0	10
treatment	0.31	0.0	0	0.00	1.00	0	1
censor	0.09	0.0	0	0.00	0.00	0	1
event	0.92	1.0	0	1.00	1.00	0	1

A FOCUS ON SURVIVAL TIME

- We expect `surv_time` to follow an exponential distribution

```r
ggstatsplot::gghistostats(data_cleaned, surv_time, binwidth = 70, title = "Histogram of surv_time")
```



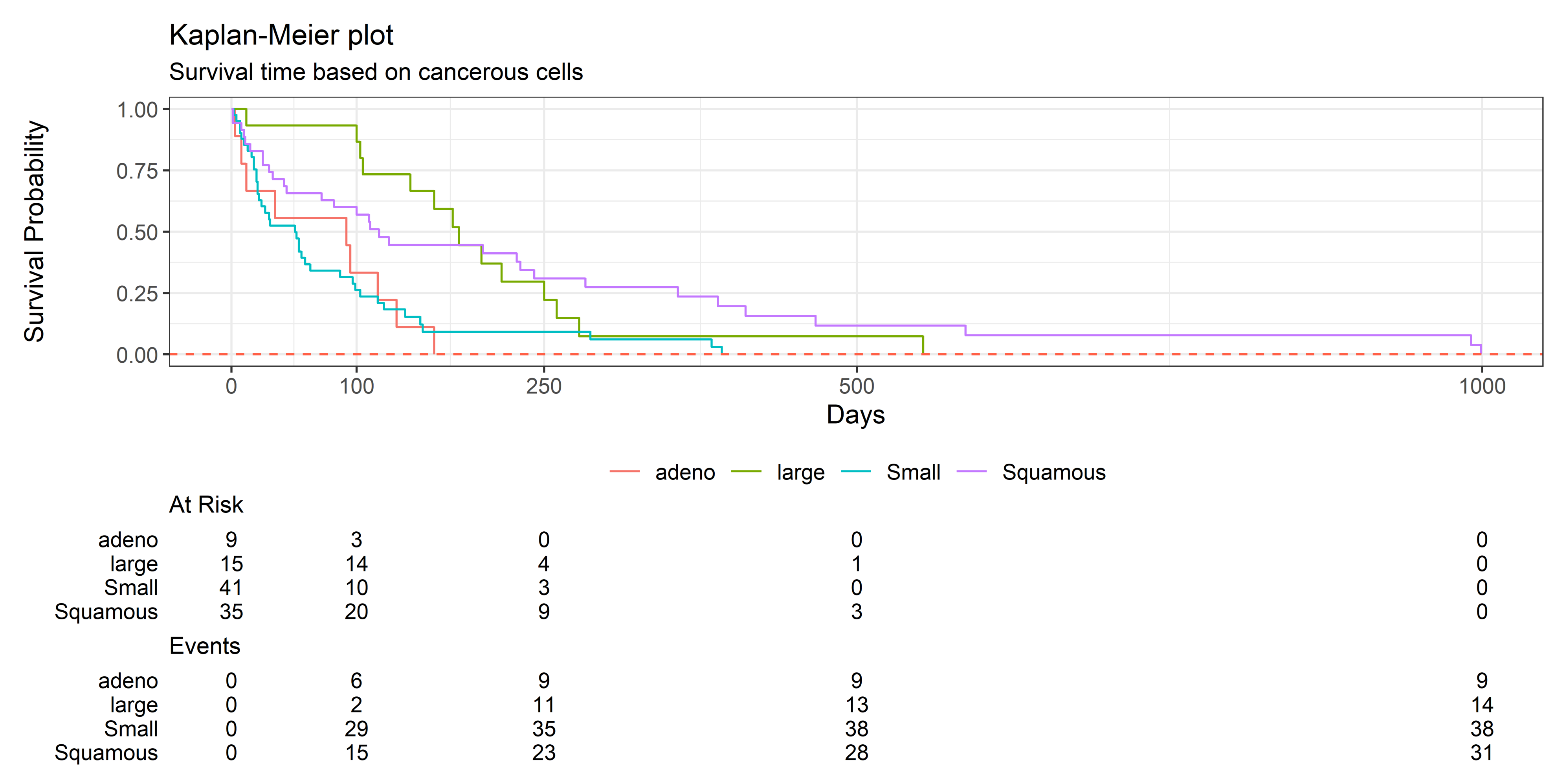
SURVIVAL ANALYSIS

Perform a survival analysis to assess the survival time? based on the cancerous cells? Consider applying survival functions/kaplan meier quartiles/cumulative incidence function/cox regression etc.

KAPLAN-MEIER PLOT

```r
survfit2(Surv(surv_time, event) ~ cell, data = data_cleaned) |>
  ggsurvfit(type = "survival") +
  labs(
    title = "Kaplan-Meier plot",
    subtitle = "Survival time based on cancerous cells",
    x = "Days"
  ) +
  scale_x_continuous(breaks = c(0, 100, 250, 500, 1000)) +
  add_risktable(times = c(0, 100, 250, 500, 1000)) +
  geom_hline(yintercept = 0, color = "tomato", linetype = "dashed")
```

KAPLAN-MEIER PLOT



Kaplan-Meier plot

Survival time based on cancerous cells

At Risk

	0	100	250	500	1000
adeno	9	3	0	0	0
large	15	14	4	1	0
Small	41	10	3	0	0
Squamous	35	20	9	3	0

Events

	0	100	250	500	1000
adeno	0	6	9	9	9
large	0	2	11	13	14
Small	0	29	35	38	38
Squamous	0	15	23	28	31

COMPARISON OF SURVIVAL CURVES

- Calculate the Log Rank Test

```
survdiff(Surv(surv_time, event) ~ cell, data = data_cleaned)
```

```
Call:
survdiff(formula = Surv(surv_time, event) ~ cell, data = data_cleaned)

               N Observed Expected (O-E)^2/E (O-E)^2/V
cell=adeno     9        9      5.2      2.78      3.04
cell=large    15       14     19.8      1.70      2.26
cell=Small    41       38     23.8      8.54     12.44
cell=Squamous 35       31     43.2      3.46      7.27

 Chisq= 18.4  on 3 degrees of freedom, p= 4e-04
```

COX REGRESSION

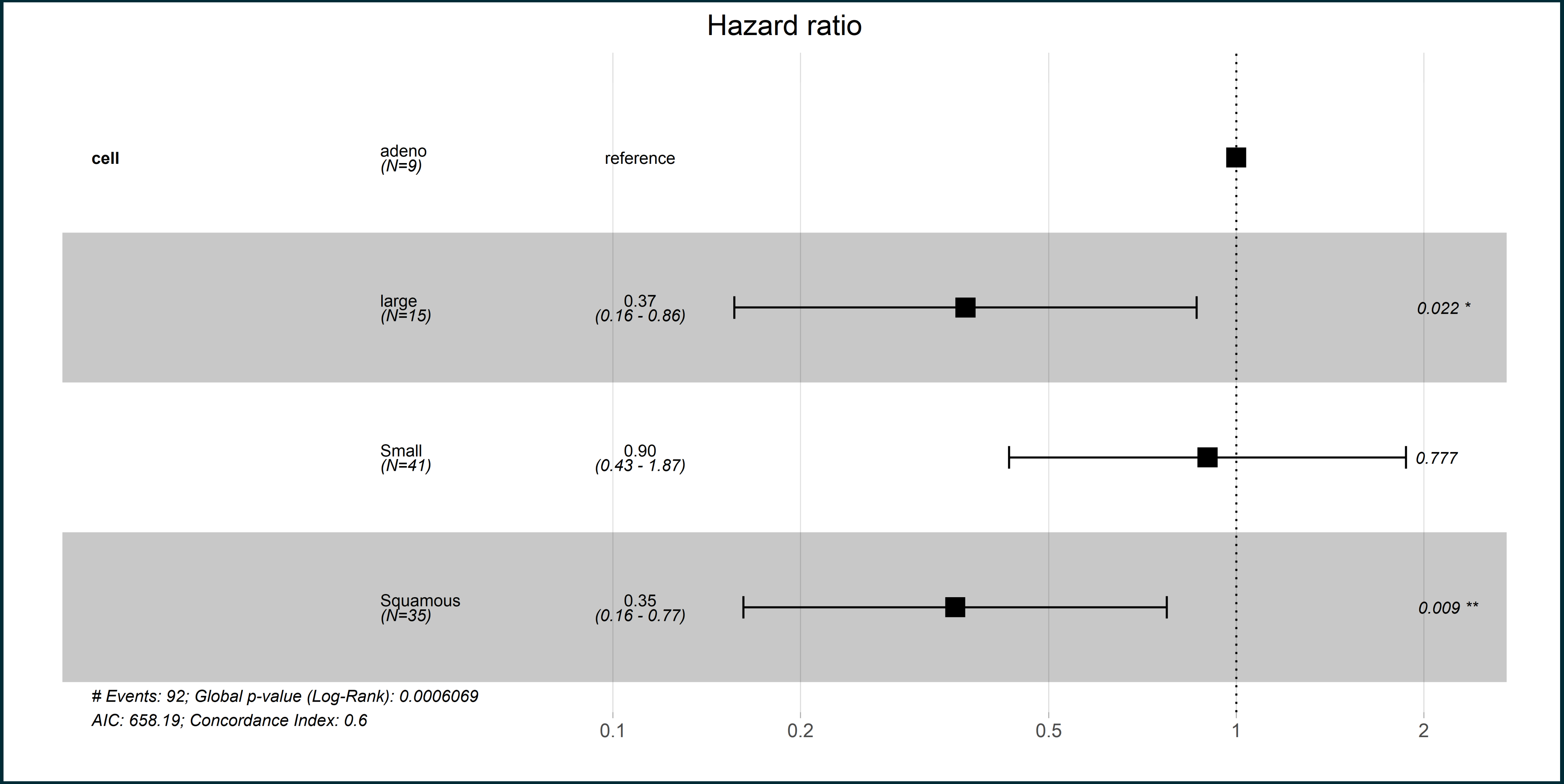
- adeno is the reference category, so the hazard ratios are relative to it

```r
cox_model <- coxph(Surv(surv_time, event) ~ cell, data = data_cleaned)
summary(cox_model)
```

```
Call:
coxph(formula = Surv(surv_time, event) ~ cell, data = data_cleaned)

  n= 100, number of events= 92 

                coef exp(coef) se(coef)      z Pr(>|z|)   
celllarge    -1.0003    0.3678   0.4358 -2.295  0.02172 * 
cellSmall    -0.1062    0.8993   0.3741 -0.284  0.77653   
cellSquamous -1.0385    0.3540   0.3990 -2.603  0.00925 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
celllarge       0.3678      2.719    0.1565    0.8640
cellSmall       0.8993      1.112    0.4320    1.8721
cellSquamous    0.3540      2.825    0.1619    0.7738

Concordance= 0.604  (se = 0.033 )
Likelihood ratio test= 17.32  on 3 df,   p=6e-04
Wald test            = 17.33  on 3 df,   p=6e-04
```

VISUALIZE THE COEFFICIENTS

```r
survminer::ggforest(cox_model)
```



MULTIVARIABLE ANALYSIS

Perform an appropriate multivariable analysis to analyze the effect of independent variables age on the hazard ratio between the different cancerous cells (var Cell)?

COX REGRESSION WITH `cell` AND `age`

```r
multivariable_cox_model <- coxph(Surv(surv_time, event) ~ cell + age, data = data_cleaned)

multivariable_cox_model |>
  tbl_regression(exponentiate = T) |>
  add_global_p() |>
  add_n(location = "level") |>
  add_nevent(location = "level") |>
  bold_labels() |>
  bold_p() |>
  italicize_levels()
```

Characteristic	N	Event N	HR	95% CI	p-value
cell					**<0.001**
adeno	9	9	—	—	
large	15	14	0.36	0.15, 0.85	
Small	41	38	0.85	0.40, 1.80	
Squamous	35	31	0.34	0.15, 0.74	
age	100	92	1.01	0.99, 1.03	0.5

Abbreviations: CI = Confidence Interval, HR = Hazard Ratio

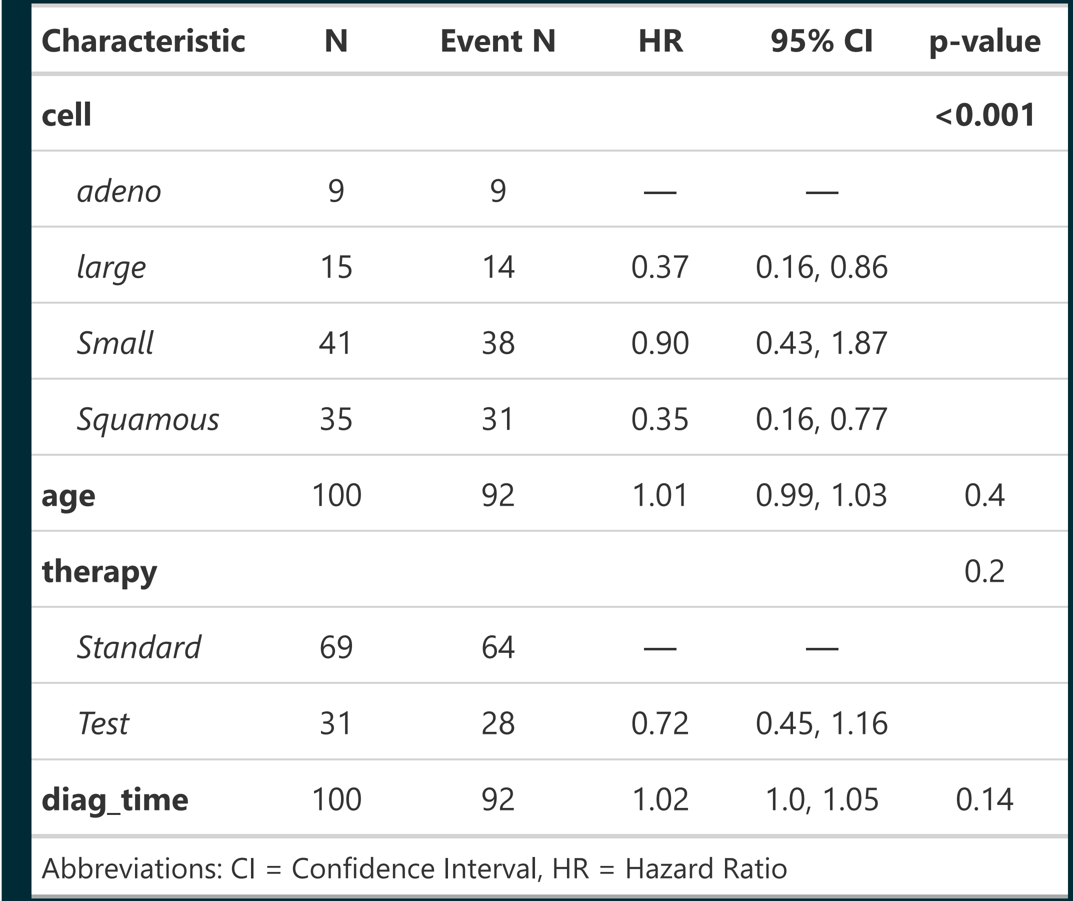
DIAGNOSTIC OF THE MODEL

- Testing the proportional hazards assumption for the multivariable Cox regression model

```r
survminer::ggcoxzph(cox.zph(multivariable_cox_model))
```

DEEPER ANALYSIS

```r
data_cleaned |>
  select(surv_time, event, cell, age,
         therapy, diag_time) |>
  tbl_uvregression(
    method = coxph,
    y = Surv(surv_time, event),
    exponentiate = T
  ) |>
  add_global_p() |>
  add_n(location = "level") |>
  add_nevent(location = "level") |>
  bold_labels() |>
  bold_p() |>
  italicize_levels()
```



Characteristic	N	Event N	HR	95% CI	p-value
cell					**<0.001**
adeno	9	9	—	—	
large	15	14	0.37	0.16, 0.86	
Small	41	38	0.90	0.43, 1.87	
Squamous	35	31	0.35	0.16, 0.77	
age	100	92	1.01	0.99, 1.03	0.4
therapy					0.2
Standard	69	64	—	—	
Test	31	28	0.72	0.45, 1.16	
diag_time	100	92	1.02	1.0, 1.05	0.14

Abbreviations: CI = Confidence Interval, HR = Hazard Ratio

JUST A CHECK

- `kps` (Key Performance Status) was not included because it’s a measure of the patients that we observe during the therapy not before, as `age`, `diag_time` or `cell`
- If included, the model results are as follows:

```r
coxph(Surv(surv_time, event) ~ kps, data = data_cleaned)
```

```
Call:
coxph(formula = Surv(surv_time, event) ~ kps, data = data_cleaned)

         coef exp(coef)  se(coef)     z        p
kps -0.029414  0.971014  0.006193 -4.75 2.04e-06

Likelihood ratio test=22.31  on 1 df, p=2.319e-06
n= 100, number of events= 92
```

WHAT ABOUT THE THERAPY PURPOSED?

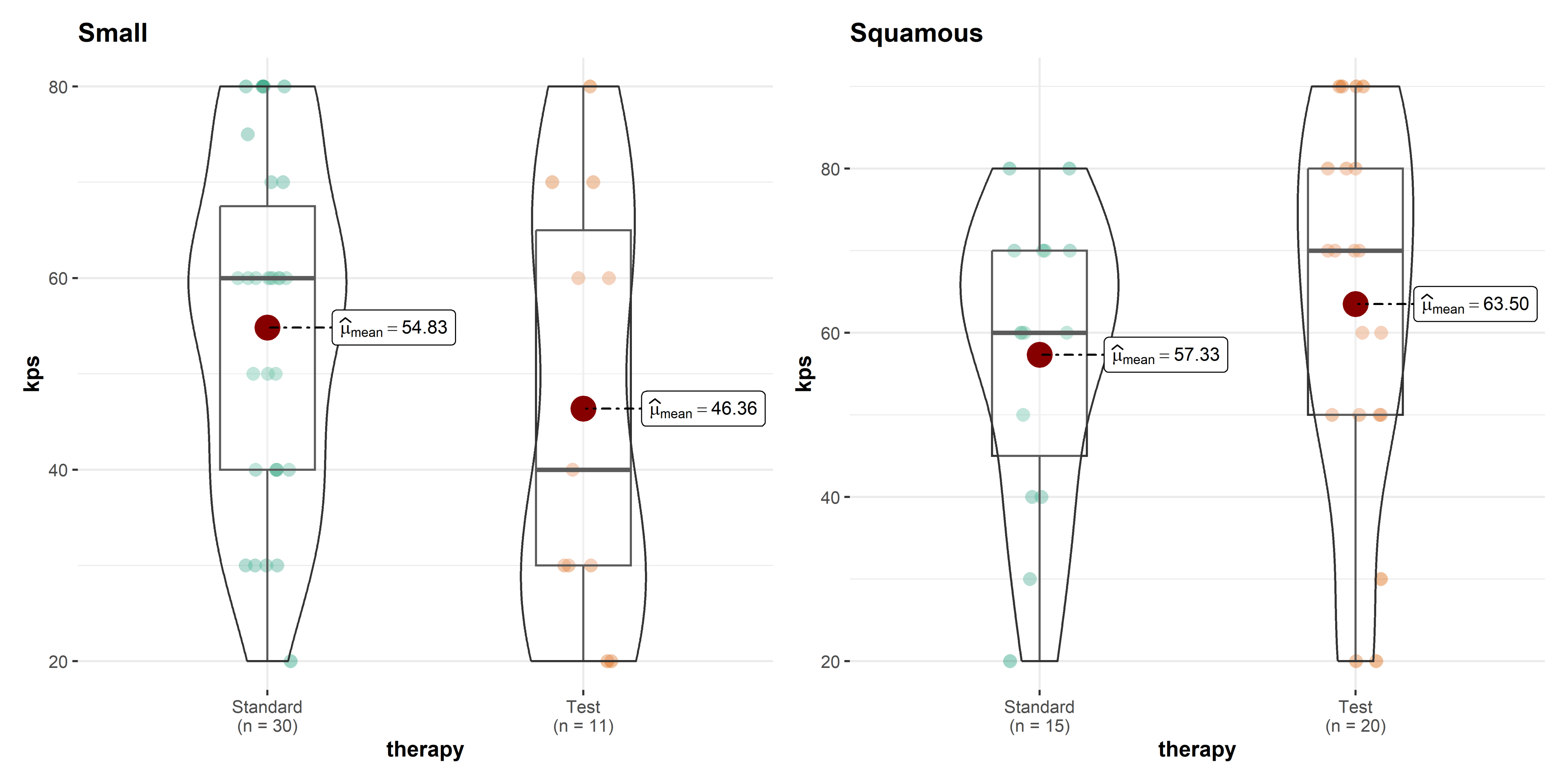
```r
data_cleaned |>
  semi_join(data_cleaned |> filter(therapy == "Test"), by = "cell") |>
  select(therapy, event, surv_time, cell, kps, diag_time) |>
  tbl_strata(
    strata = cell,
    .tbl_fun =
      \(x) x |>
      tbl_summary(
        by = therapy,
        type = kps ~ "continuous"
      ) |>
      add_p()
  )
```

	Small			Squamous		
Characteristic	**Standard** N = 30[1]	**Test** N = 11[1]	**p-value**[2]	**Standard** N = 15[1]	**Test** N = 20[1]	**p-value**[2]
event	28 (93%)	10 (91%)	>0.9	13 (87%)	18 (90%)	>0.9
surv_time	53 (20, 122)	21 (8, 87)	0.080	100 (25, 144)	157 (32, 373)	0.3
kps	60 (40, 70)	40 (30, 70)	0.2	60 (40, 70)	70 (50, 80)	0.3
diag_time	4 (3, 11)	4 (2, 11)	>0.9	9 (5, 11)	7 (3, 13)	0.5

[1] n (%); Median (Q1, Q3)

[2] Fisher's exact test; Wilcoxon rank sum test

LET'S PLOT IT!


Small
80
60
40
20
kps
$\hat{\mu}_{\text{mean}} = 54.83$
$\hat{\mu}_{\text{mean}} = 46.36$
Standard
(n = 30)
Test
(n = 11)
therapy
Squamous
80
60
40
20
kps
$\hat{\mu}_{\text{mean}} = 57.33$
$\hat{\mu}_{\text{mean}} = 63.50$
Standard
(n = 15)
Test
(n = 20)
therapy

THANKS