

# Survival analysis-I

Shivangi Agarwal

PhD

# WHAT WE WILL LEARN

Introduction  
to Survival  
analysis

Calculation of  
survival time  
for an  
example  
dataset

Methods:  
Kaplan Meier  
and COX PH

Kaplan Meier  
plots

Log rank test:  
Estimation of  
chi square

Survival &  
Hazard  
function

COX PH  
model

Hazard ratio

Hands on KM  
plots: lungs  
dataset

Hands on KM  
plots: TCGA  
dataset

Hands on  
COX PH: lungs  
dataset

Hands on  
COX PH:  
TCGA dataset

# SURVIVAL ANALYSIS

- Survival analysis deals with the prediction of events at a specified time.
- It deals with the occurrence of an interesting event within a specified time and failure of it produces censored observations i.e incomplete observations.
- Generally, survival analysis lets you model the time until an event occurs, or compare the time-to-event between different groups, or how time-to-event correlates with quantitative variables.
- The survival function, is the probability an individual survives (or, the probability that the event of interest does not occur) up to and including time  $t$ .

# EVENTS

Relapse of  
a disease

Death

Progression

## ***Time to event :***

The time from 'response to treatment' (complete remission) to the occurrence of the event of interest.

The two most important measures in cancer studies include:

- i) the time to death; and
- ii) the relapse-free survival time, which corresponds to the time between response to treatment and recurrence of the disease. It's also known as disease-free survival time and event-free survival time.

# EXAMPLES FROM CANCER

- ☐ Time from surgery to death
- ☐ Time from start of treatment to progression
- ☐ Time from response to recurrence

Some other examples include:

- ☐ Time from HIV infection to development of AIDS
- ☐ Time to heart attack
- ☐ Time to onset of substance abuse
- ☐ Time to initiation of sexual activity
- ☐ Time to machine malfunction

# APPLICATIONS

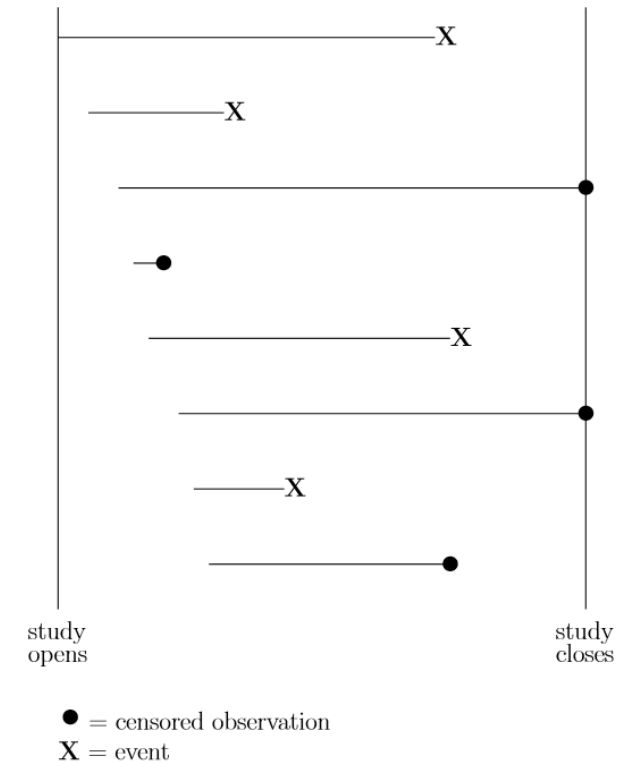
Survival analysis is used in a variety of field such as:

- Cancer studies for patients survival time analyses,
  - ✓ What is the impact of certain clinical characteristics on patient's survival
  - ✓ What is the probability that an individual survives 3 years?
  - ✓ Are there differences in survival between groups of patients?
  - ✓ To estimate probability of not experiencing event of interest (not dying = “surviving”) over any given time period (e.g. 5 year survival rate).
- Sociology for “event-history analysis”,
- In engineering for “failure-time analysis”

# CENSORING

- Event of interest not observed for all individuals.
- Fixed censoring: event has not occurred when study has ended or data analysis is performed.
- Loss to follow-up: individual has been lost to follow-up (e.g. he/she no longer wishes to take part in study).
- Type I censoring: We only have the survival information up to a fixed time.
- Type II censoring: We only observe the first  $r$  smallest survival times.
- Random censoring
- Interval censoring

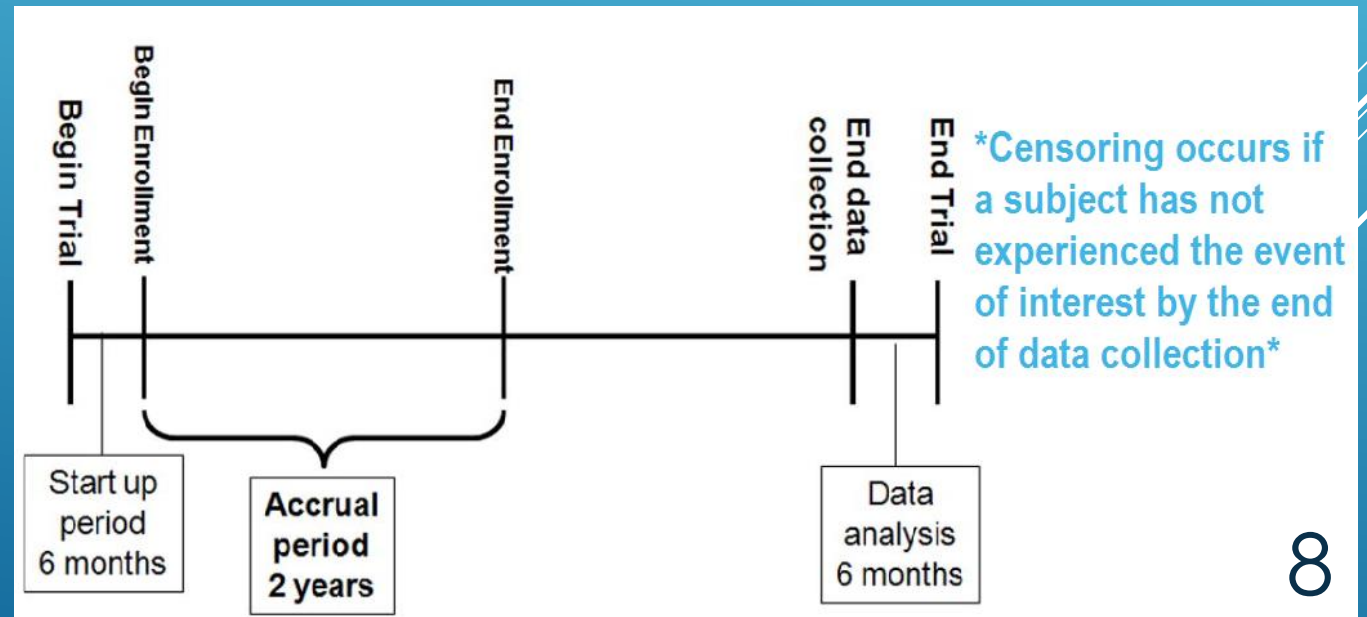
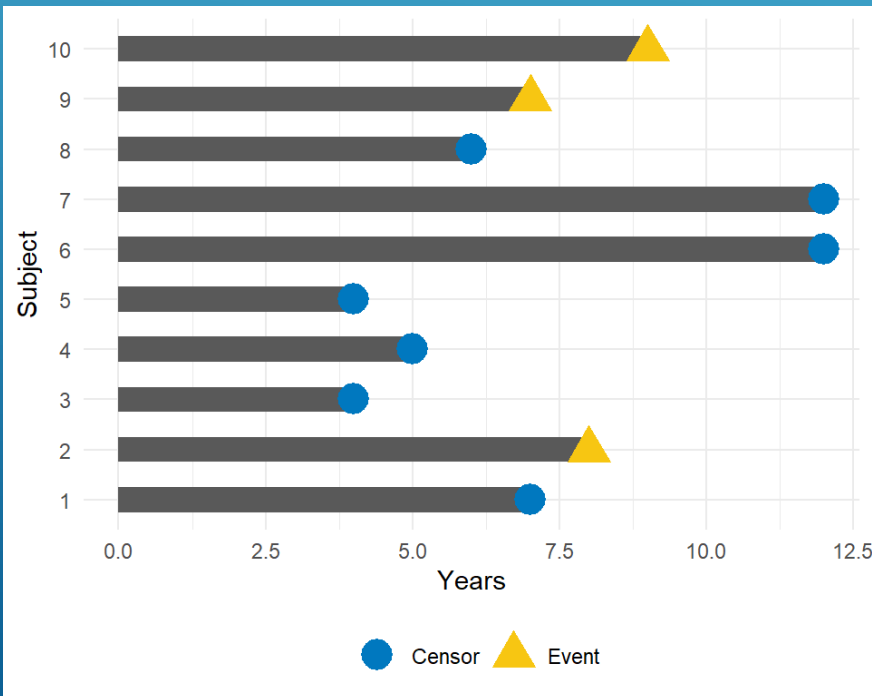
Illustration of survival data



# CENSORING

Censoring may arise in the following ways:

- a patient has not (yet) experienced the event of interest, such as relapse or death, within the study time period;
- a patient is lost to follow-up during the study period;
- a patient experiences a different event that makes further follow-up impossible.
- Censoring is a type of missing data problem unique to survival analysis.
- The sample is censored in that you only know that the individual survived up to the loss to followup, but you don't know anything about survival after that.





# Example of time to event data

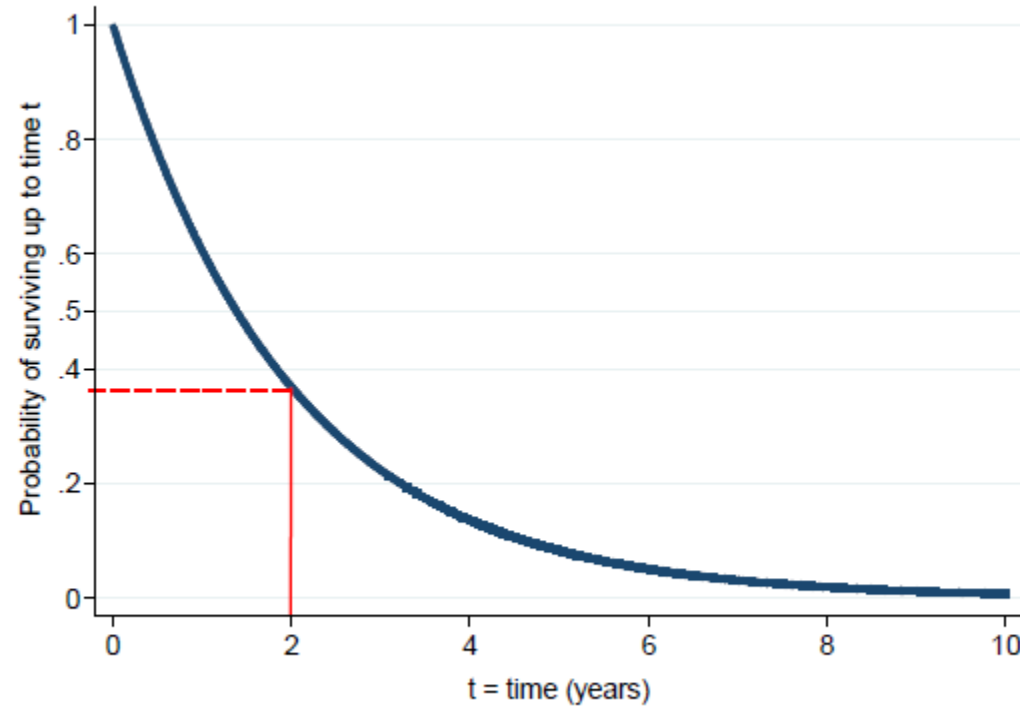
Weeks to death or censoring (\*) in 20 adults with recurrent astrocytoma:

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

ID	death	weeks
1	1	6
2	1	13
3	1	21
4	1	30
5	0	31
6	1	37
7	1	38
8	0	47
9	1	49
10	1	50
11	1	63
12	1	79
13	0	80
14	0	82
15	0	82
16	1	86
17	1	98
18	0	149
19	1	202
20	1	219

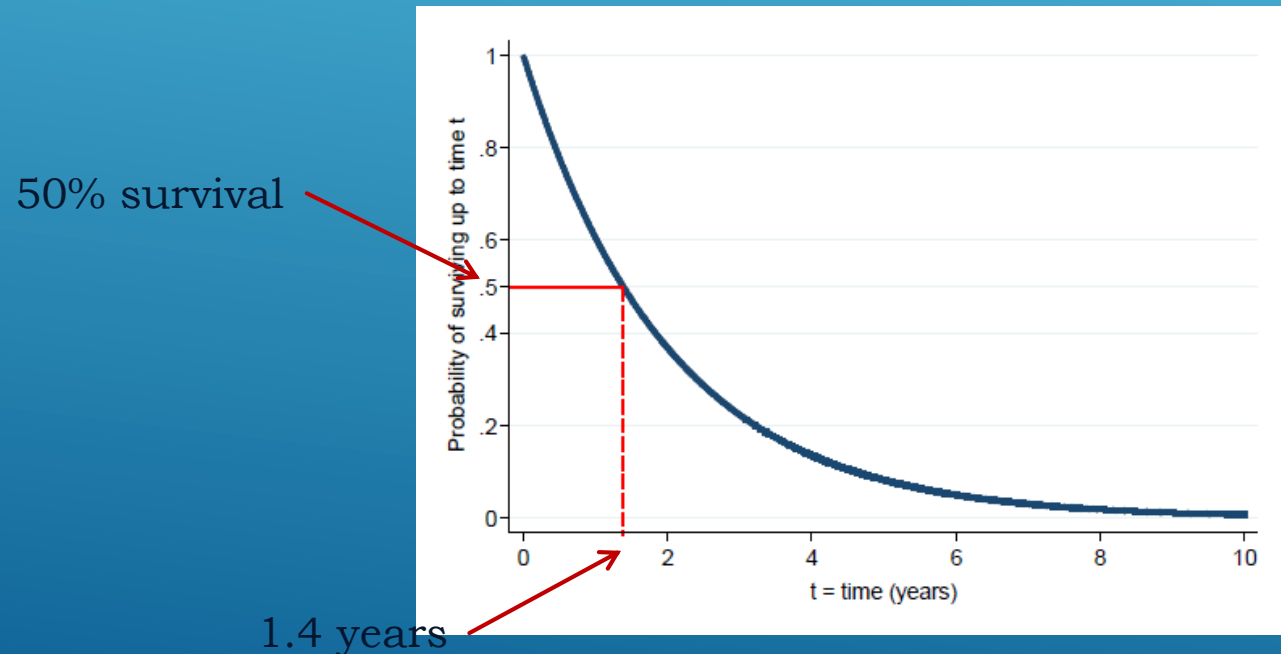
# Estimating a survival rate

- Probability of surviving up to 2 years = 0.37.



# Median survival time

- It is the time (expressed in months or years) when half the patients are expected to be alive. It means that the chance of surviving beyond that time is 50%.
- Median survival time = 1.4 years, since the probability of surviving up to 1.4 years is 0.5.



# METHODS OF SURVIVAL ANALYSIS

**Kaplan-Meier method**

**Cox Proportional hazard model**

# KAPLAN MEIER METHOD

- It's a non-parametric statistic that allows us to estimate the survival function and thus not based on underlying probability distribution.
- The curve is horizontal over periods where no event occurs, then drops vertically corresponding to a change in the survival function at each time an event occurs.
- The Kaplan-Meier method is used in survival distribution using the Kaplan-Meier estimator for truncated or censored data.
- The Kaplan–Meier estimates are based on the number of patients (each patient as a row of data) from the total number of patients who survive for a certain time after treatment. (which is the event).

$$S(t) = P(T < t) \times P(T > t)$$

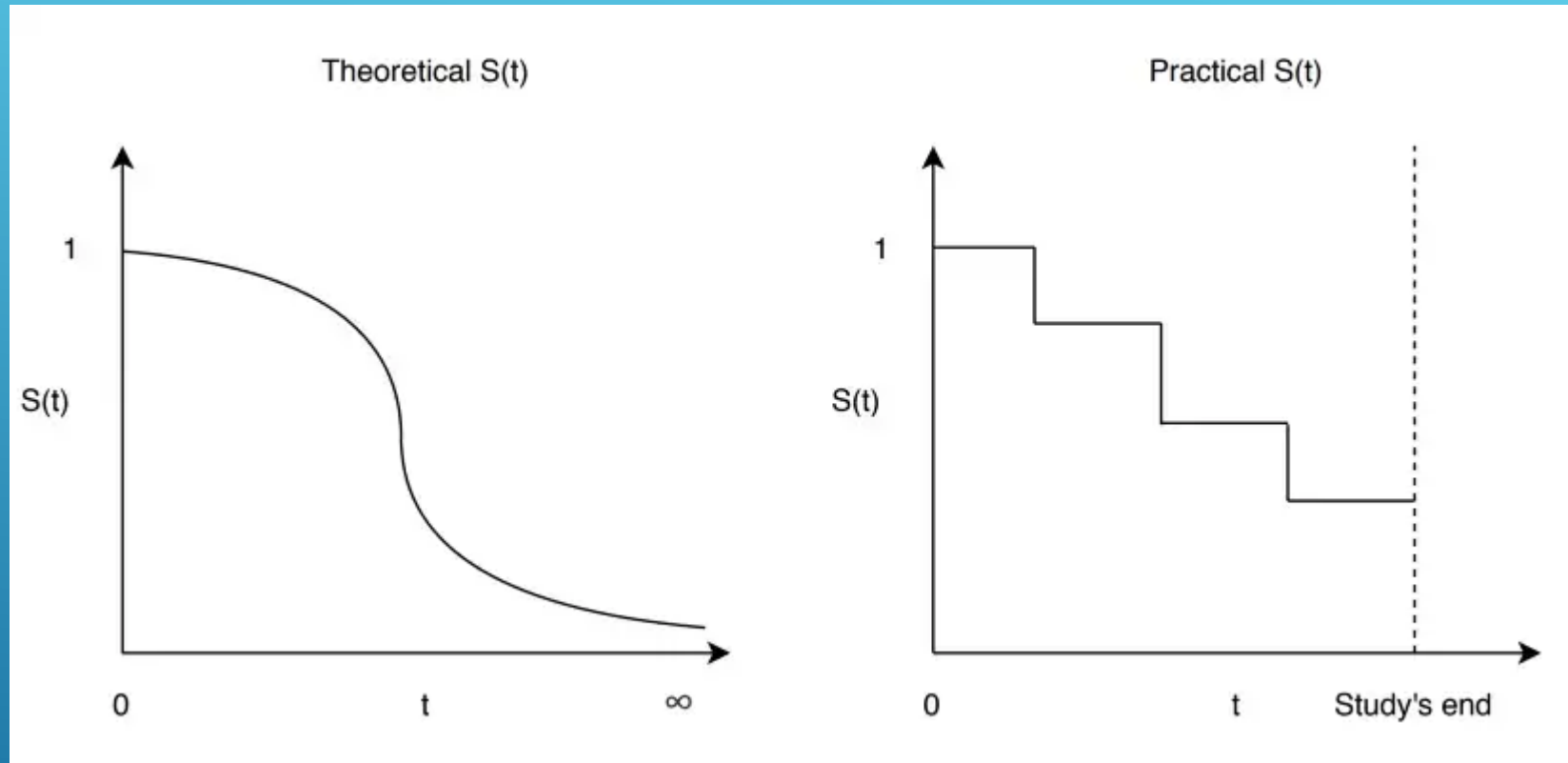
T is the time of death,

$P(T < t)$  is the probability of survival before time t

$P(T > t)$  is the probability that the time of death is greater than some time t.

S is a probability, so  $0 \leq S(t) \leq 1$ , since survival times are always positive ( $T \geq 0$ ).

# SURVIVAL CURVES



Each drop in the survival function (approximated by the Kaplan-Meier estimator) is caused by the event of interest happening for at least one observation.

# Kaplan-Meier (KM) estimation of survivor function

ID	Time in weeks	Death (event)
P1	6	1
P2	13	1
P3	21	1
P4	30	1
P5	31+	0
P6	37	1
P7	38	1
P8	47+	0
P9	49	1
P10	50	1
P11	63	1
P12	79	1
P13	80+	0
P14	82+	0
P15	82+	0
P16	86	1
P17	98	1
P18	149+	0
P19	202	1
P20	219	1

20 adults with recurrent  
astrocytoma

# Kaplan-Meier (KM) estimation of survivor function at first event

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

Time (weeks)	No of patients at risk (r)	No of deaths at time t (d)	Prob of death (d/r)	1-d/r	$S(t)=P(T<t) \times P(T>t)$
0	20	0	0	1	1
6	20	1 (1st event)	$1/20=0.05$	$1-1/20=0.95$	$1 \times 0.95=0.95$

- **20** individuals in study at  $t=0$ .
- First death at  $t=6$  weeks.
- No individuals censored before  $t=6$ .
- Probability of death for each individual:  **$1/20=0.05$**
- Therefore probability of surviving beyond  $t=6$  is  **$(1-0.05)=0.95=19/20$** .



# Kaplan-Meier (KM) estimation of survivor function at second event

	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

Time (weeks)	No of patients at risk (r)	No of deaths at time t (d)	Prob of death (d/r)	1-d/r	S(t)=P(T<t)xP(T>t)
6	20	1 (1st event)	1/20=0.05	1-1/20=0.95	1x0.95=0.95
13	19	1 (2 <sup>nd</sup> event)	1/19=0.053	1-1/19=0.947	0.95x0.947=0.90

- **19** individuals in study between t=6 and t=13.
- Second death at t=13.
- No individuals censored between t=6 and t=13.
- Probability of death for each individual: **1/19=0.053**
- Therefore probability of surviving beyond t=13 is **0.95 x 0.947 =0.90**.
  - with **0.95=(1-(1/20))** and **0.947=(1-(1/19))**

# Kaplan-Meier (KM) estimation of survivor function at third and fourth event

		21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

Time (weeks)	No of patients at risk (r)	No of deaths at time t (d)	Prob of death (d/r)	1-d/r	S(t)=P(T<t)xP(T>t)
6	20	1 (1st event)	1/20=0.05	1-1/20=0.95	1x0.95=0.95
13	19	1 (2 <sup>nd</sup> event)	1/19=0.053	1-1/19=0.947	0.95x0.947=0.90
21	18	1 (3 <sup>rd</sup> event)	1/18=0.056	1-1/18=0.944	0.90x0.944=0.85
30	17	1 (4 <sup>th</sup> event)	1/17=0.059	1-1/17=0.941	0.85x0.941=0.80

- **18** individuals in study between t=13 and t=21.
- Probability of death for each individual: **1/18=0.056**
- Probability of surviving beyond t=21 is **0.90 x (1-(1/18)) =0.85**.
- **17** individuals in study between t=21 and t=30.
- Probability of death for each individual: **1/17=0.059**
- Probability of surviving beyond t=30 is **0.85 x (1-(1/17)) =0.80**.

# Kaplan-Meier (KM) estimation of survivor function at fifth and sixth event

				31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

Time (weeks)	No of patients at risk (r)	No of deaths at time t (d)	Prob of death (d/r)	1-d/r	$S(t)=P(T<t) \times P(T>t)$
6	20	1 (1st event)	$1/20=0.05$	$1-1/20=0.95$	$1 \times 0.95=0.95$
13	19	1 (2 <sup>nd</sup> event)	$1/19=0.053$	$1-1/19=0.947$	$0.95 \times 0.947=0.90$
21	18	1 (3 <sup>rd</sup> event)	$1/18=0.056$	$1-1/18=0.944$	$0.90 \times 0.944=0.85$
30	17	1 (4 <sup>th</sup> event)	$1/17=0.059$	$1-1/17=0.941$	$0.85 \times 0.941=0.80$
31	16	0 (no event ?)	0	1	$0.80 \times 1=0.80$
37	15	1 (6 <sup>th</sup> event)	$1/15=0.067$	$1-1/15=0.933$	$0.80 \times 0.933=0.747$

- 16 individuals in study between  $t=30$  and  $t=31$ .
- 1 individual censored at  $t=31$ .
- **Probability of surviving beyond  $t=31$  remains at 0.80.**
- 15 individuals in study between  $t=31$  and  $t=37$ .
- Probability of surviving beyond  $t=37$  is  **$0.80 \times (1-(1/15)) = 0.747$ .**

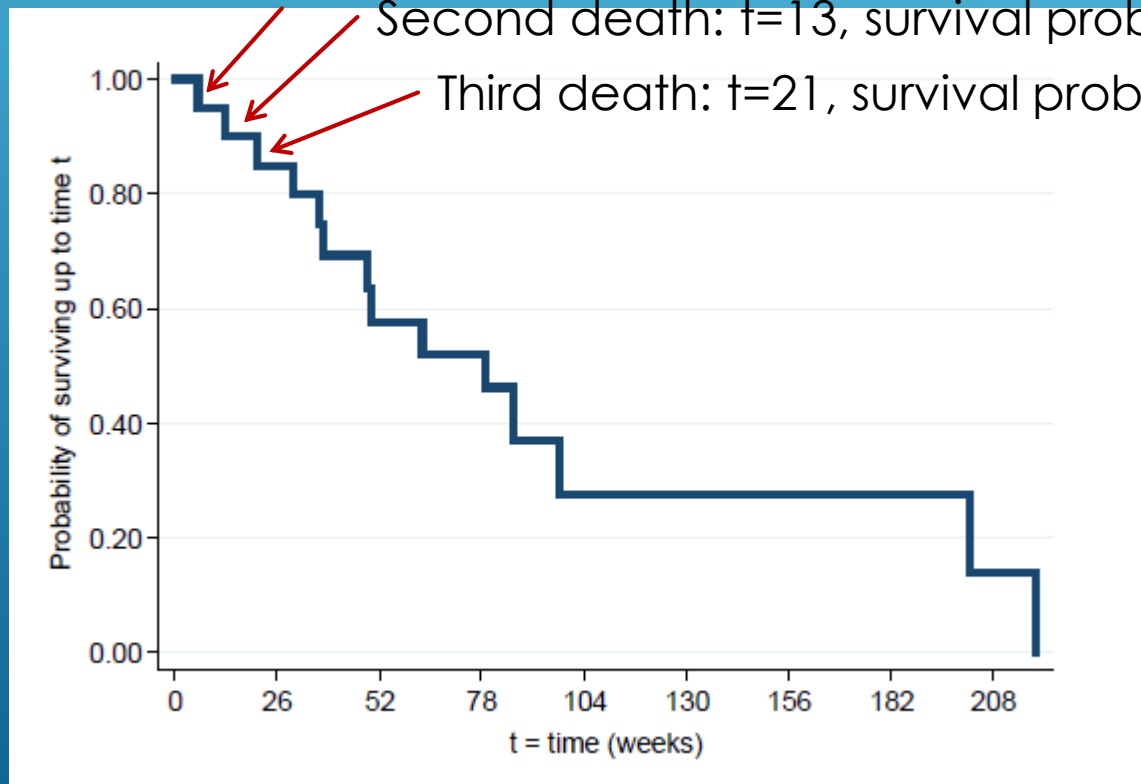
# Kaplan-Meier (KM) plot of survivor function

- Continue these calculations until reaching the longest event time.
- K-M plot drawn as a step function:

First death:  $t=6$ , survival probability=0.95

Second death:  $t=13$ , survival probability=0.90

Third death:  $t=21$ , survival probability=0.85



# COMPARING SURVIVAL BETWEEN TWO GROUPS

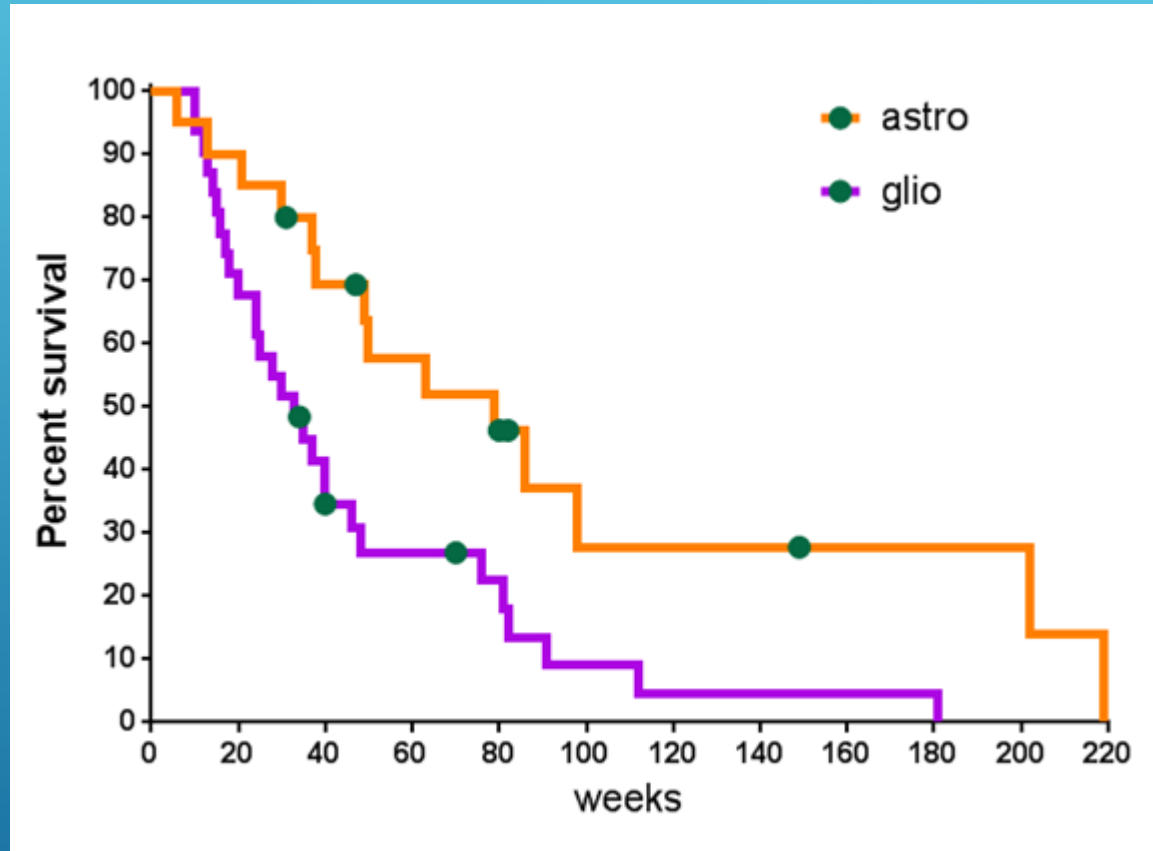
- Weeks to death or censoring (\*) in **20 adults** with recurrent astrocytoma:

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- Weeks to death or censoring (\*) in **31 adults** with recurrent glioblastoma:

10	10	12	13	14	15	16	17	18	20
24	24	25	28	30	33	34*	35	37	40
40	40*	46	48	70*	76	81	82	91	112
181									

# KM Plot of survivor function by tumor type



- Survival chances appear better in individuals with astrocytoma than with glioblastoma, but is the **difference between groups statistically significant?**

# LOG RANK TEST

**Log rank test:** tests null hypothesis of no difference between samples in probability of an event (death in this example) at any time point during follow-up.

The log rank test is a non-parametric test, which makes no assumptions about the survival distributions. Essentially, the log rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true (i.e., if the survival curves were identical). The log rank statistic is approximately distributed as a chi-square test statistic.

Additionally, there is the **proportional hazards assumption** — the hazard ratio should be constant throughout the study period. In practice, this means that the log-rank test might not be an appropriate test if the survival curves cross.

# LOG RANK TEST BETWEEN TWO GROUPS

## Study: Clinical trials of two cancer drugs

Time (years)	No. of patients at risk (n1)	No of deaths (d1)	No. of patients at risk (n2)	No of deaths (d2)	N (n1+n2)	D (d1+d2)	Expected events in G1 (n1*D/N)	Expected events in G2 (n2*D/N)
0	18	0	18	0	36	0	0	0
1	18	0	18	1	36	1	0.5	0.5
2	18	1	17	0	35	1	0.514286	0.485714
3	17	1	17	1	34	2	1	1
4	16	0	16	1	32	1	0.5	0.5
5	16	3	15	0	31	3	1.548387	1.451613
6	12	0	15	1	27	1	0.444444	0.555556
8	12	3	12	1	24	4	2	2
9	9	0	10	0	19	0	0	0
10	8	0	9	2	17	2	0.941176	1.058824
11	6	2	5	0	11	2	1.090909	0.909091
12	4	2	5	0	9	2	0.888889	1.111111
13	2	0	3	1	5	1	0.4	0.6
		12		8			9.82092	10.17191



# LOG RANK TEST BETWEEN TWO GROUPS

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$LR = \frac{(Obs_A - Exp_A)^2}{Exp_A} + \frac{(Obs_B - Exp_B)^2}{Exp_B}$$

$$LR = \frac{(12 - 9.828)^2}{9.828} + \frac{(8 - 10.172)^2}{10.172} = 0.944$$

If the null hypothesis is true (that the two survival distributions are the same), then the log-rank test statistic has a chi-square distribution with one degree of freedom, i.e.

$$LR \sim \chi^2(1)$$



P-value is 0.33

28

**So we cannot reject the null hypothesis that the survival rates for the two drugs under trial are statistically the same**

**Chi-square Distribution Table**

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

# HANDS ON

# HANDS ON COMPONENTS

Downloading clinical and gene expression data from TCGA

- ❑ Kaplan Meier on lung dataset inbuilt
- ❑ Survival analysis on TCGA breast cancer dataset
  - ✓ Survival based on race
  - ✓ Survival based on clinical stages
  - ✓ Survival based on gene expression

# RSTUDIO: ESTIMATION OF SURVIVAL

- ❖ `install.packages(c("survival", "survminer"))`
- ❖ `library("survival")`
- ❖ `library("survminer")`
- ❖ `View(lung)`
- ❖ `head(lung)`

**#The function `survfit()` [in survival package] can be used to compute kaplan-Meier survival estimate.**

- ❖ `fit <- survfit(Surv(time, status) ~ sex, data = lung)`
- ❖ `print(fit)`

**# Summary of survival curves**

- ❖ `summary(fit)`
- ❖ `res.sum <- surv_summary(fit)`
- ❖ `head(res.sum)`

**# Access to the sort summary table**

- ❖ `summary(fit)$table`
- ❖ `d <- data.frame(time = fit$time,  
                  n.risk = fit$n.risk,  
                  n.event = fit$n.event,  
                  n.censor = fit$n.censor,  
                  surv = fit$surv,  
                  upper = fit$upper,  
                  lower = fit$lower)`
- ❖ `head(d)`

# RSTUDIO: MAKING SURVIVAL PLOTS

```
# Change color, linetype by strata, risk.table color by strata
ggsurvplot(fit,
  pval=TRUE, conf.int = TRUE,
  risk.table.col = "strata", # Change risk table color by groups
  ggtheme = theme_bw(), # Change ggplot2 theme
  palette = c("#E7B800", "#2E9FDF"),
  xlim = c(0, 600))

ggsurvplot(fit,
  pval=TRUE, conf.int = TRUE,
  risk.table.col = "strata", # Change risk table color by groups
  ggtheme = theme_bw(), # Change ggplot2 theme
  palette = c("#E7B800", "#2E9FDF"),
  fun = "event")

ggsurvplot(fit,
  pval = TRUE, conf.int = TRUE,
  risk.table = TRUE, # Add risk table
  risk.table.col = "strata", # Change risk table color by groups
  linetype = "strata", # Change line type by groups
  surv.median.line = "hv", # Specify median survival
  ggtheme = theme_bw(), # Change ggplot2 theme
  palette = c("#E7B800", "#2E9FDF"))
```

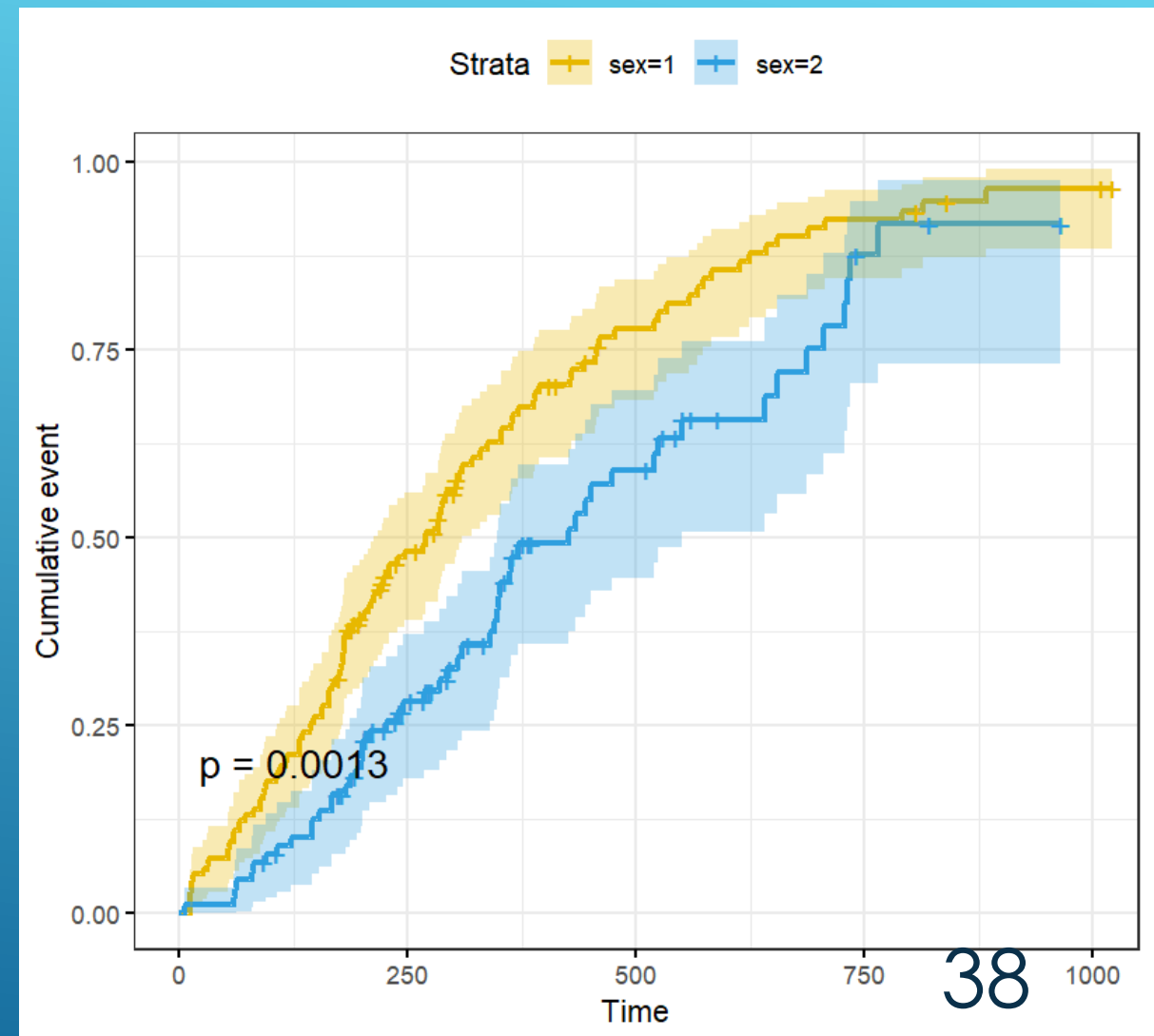
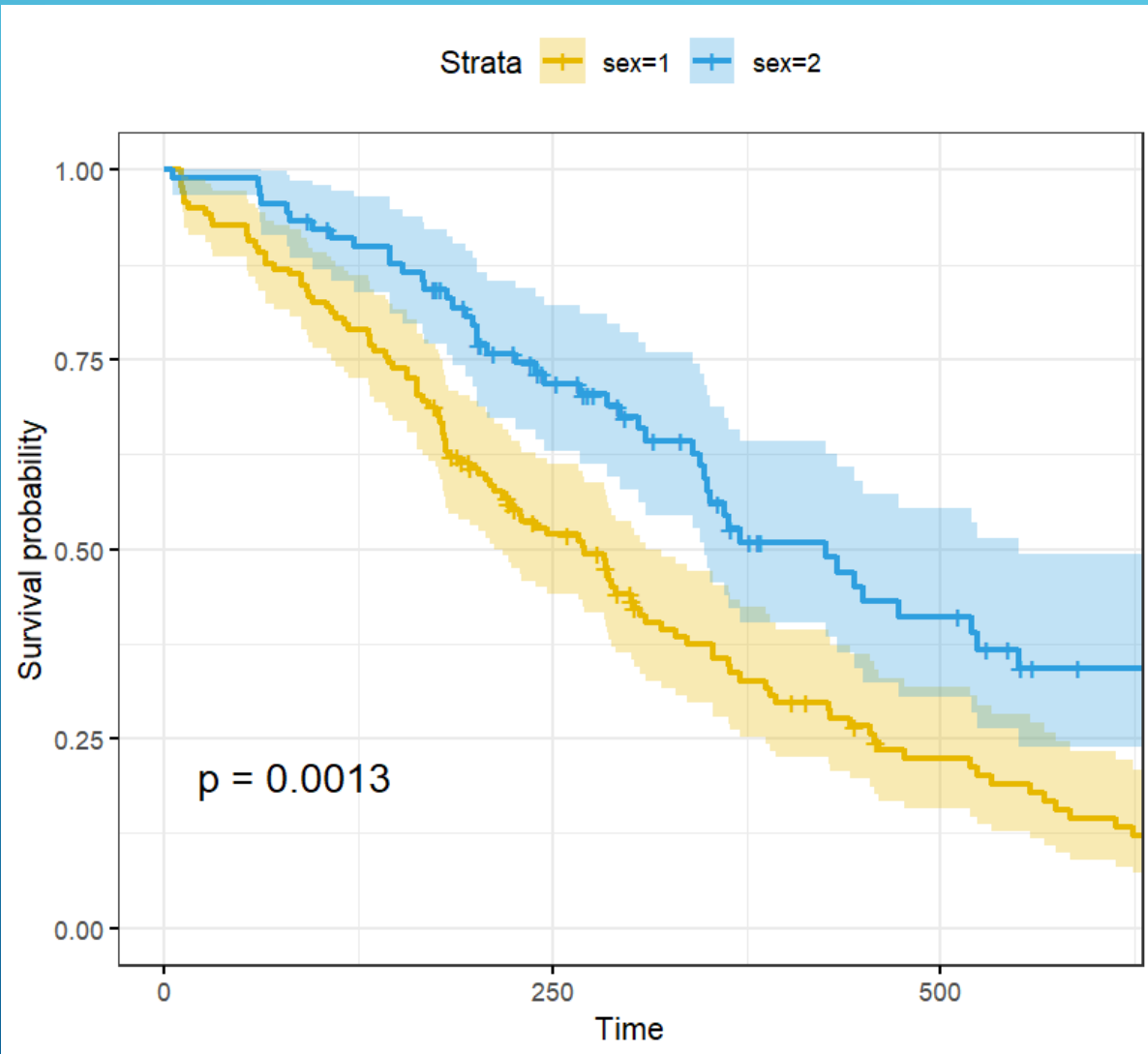
# RSTUDIO: LOG RANK TEST

**#Log-Rank test comparing survival curves: survdiff()**

**surv\_diff <- survdiff(Surv(time, status) ~ sex, data = lung)**

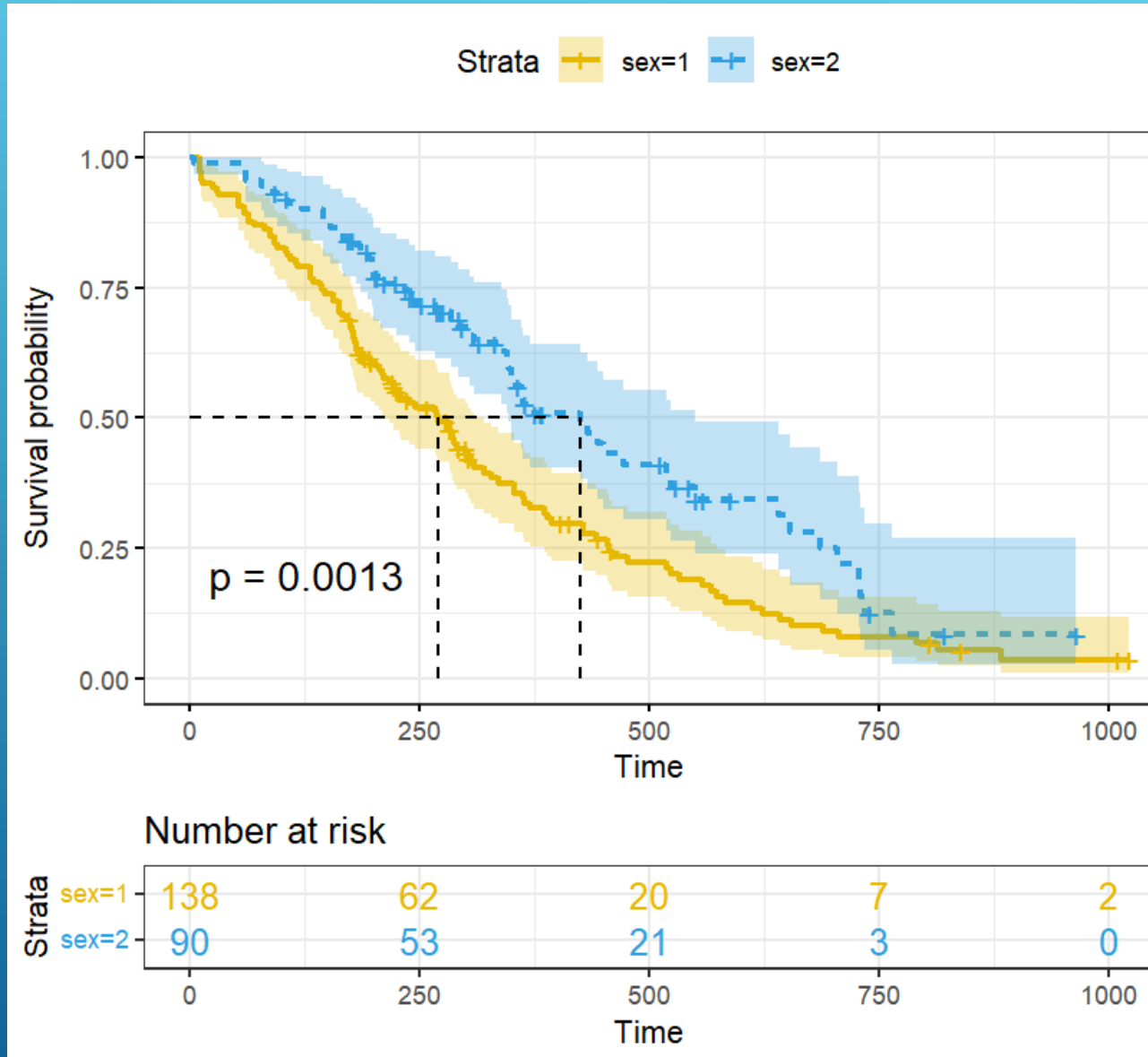
**surv\_diff**

# SURVIVAL PLOTS





# SURVIVAL PLOTS



# ESTIMATION OF SURVIVAL ON TCGA BREAST CANCER DATASET

# DATA PREPARATION

## Gene Expression data for survival analysis

Download expression data of BRCA patients using TCGA Assembler  
`geneExp2 <- DownloadRNASeqData(cancerType = "BRCA", assayPlatform  
= "gene.normalized_RNAseq", tissueType = "TP",  
saveFolderName = ".")`

Here, we will download expression data (BRCA\_\_gene.normalized\_RNAseq\_\_TP\_\_20230211101034.txt)  
for tumor samples, not for normal samples.

We got here 1095 cases.

# DATA PREPARATION

## Clinical data preparation for survival analysis

The clinical data contains clinical info for 1094 cases but they are in duplicate (only last column is different),

Secondly, column of interest in clinical data are **days\_to\_death**, **days\_to\_last\_followup** and **vital\_status**

1. In vital status column,  
replace alive with 0 (no event occurred/censored data)  
and replace dead with 1 (event occurred)

2. Overall survival for patients in which event occurred survival time will be "days\_to\_death"  
and for patients in which no event occurred, survival time will be "days\_to\_last\_followup"

Save as clinical-for-survival.tsv

Print even columns plus header using awk command as follows  
`awk 'NR==1; NR%2==0' clinical-for-survival.tsv > clinical-2.csv`

# DATA PREPARATION

**Merge expression and clinical data based on TCGA barcodes and saved as TCGA-2.csv**

**1095 rows & 20690 columns**

**##just minimizing the dataset for demo purpose**

**cat TCGA-2.csv | head -n500 > TCGA-3.csv**

**500 rows & 20690 columns**

# ESTIMATION OF SURVIVAL DIFFERENCE BETWEEN WHITE AND BLACK OR AFRICAN AMERICAN

```
TCGAMergedata <- read.csv(file = "TCGA-3.csv", sep = "\t", check.names = F, row.names = 1)
dim(TCGAMergedata)
View(TCGAMergedata)
getwd()
table(TCGAMergedata$vital_status)
table(TCGAMergedata$`Overall survival`)
table(TCGAMergedata$race)
#surv_object <- Surv(time = TCGAMergedata$`Overall survival`, event = TCGAMergedata$vital_status)
#fit <- survfit(surv_object~TCGAMergedata$gender)

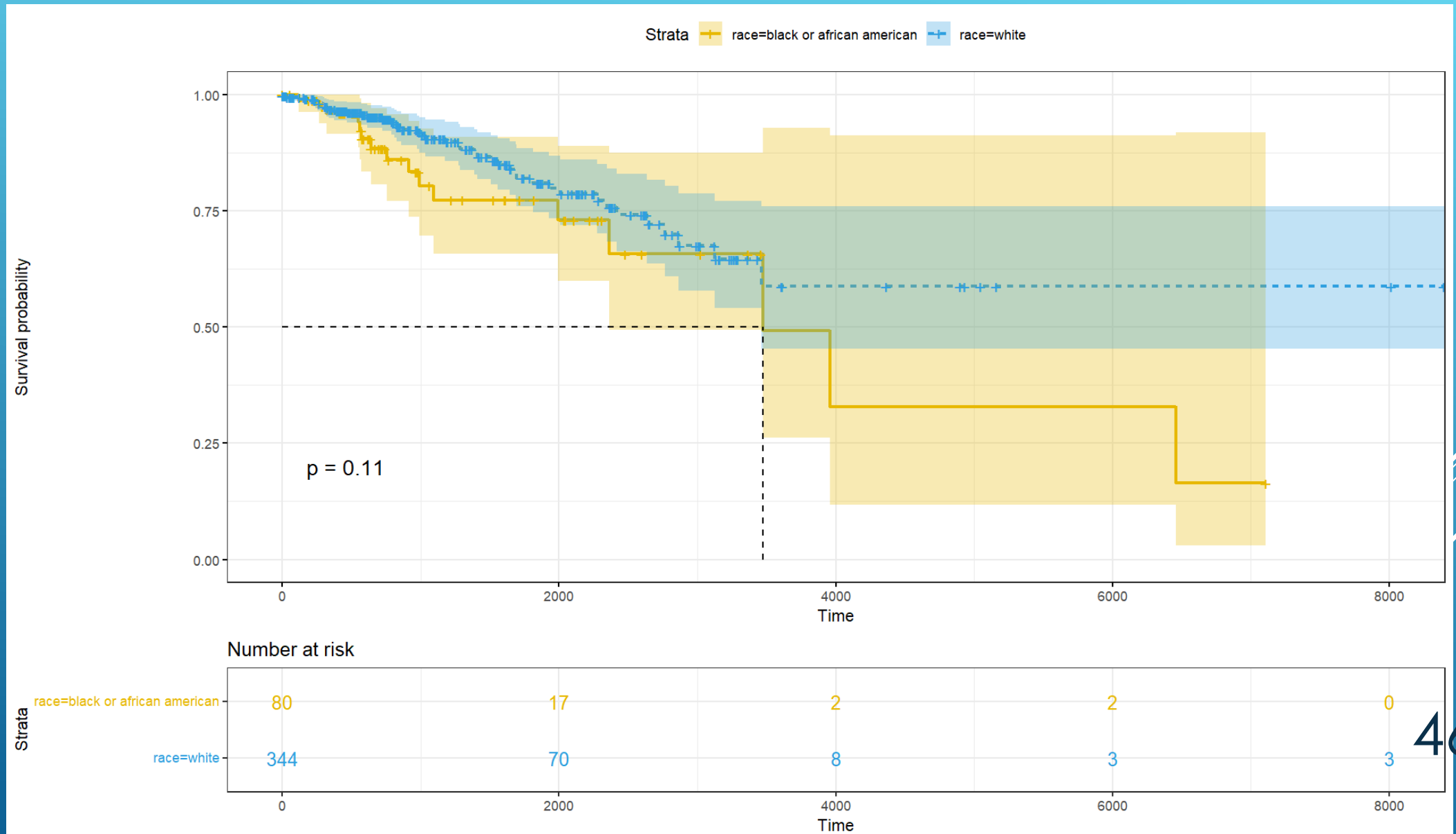
####Estimation of survival difference between white and black or African american
dataRace=subset(TCGAMergedata, (TCGAMergedata$race!='not reported')&(TCGAMergedata$race!='american
indian or alaska native')&(TCGAMergedata$race!='asian'))
table(dataRace$race)
surv_object <- Surv(time = dataRace$`Overall survival`, event = dataRace$vital_status)
fit <- survfit(surv_object~dataRace$race)
ggsurvplot(fit, data=dataRace, pval=TRUE)
surv_diff1 <- survdiff(Surv(dataRace$`Overall survival`, dataRace$vital_status) ~ dataRace$race, data =
dataRace)
surv_diff1
```

# ESTIMATION OF SURVIVAL DIFFERENCE BETWEEN WHITE AND BLACK OR AFRICAN AMERICAN

```
ggsurvplot(fit,data= dataRace,  
  pval=TRUE, conf.int = TRUE,  
  risk.table.col = "strata", # Change risk table color by groups  
  ggtheme = theme_bw(), # Change ggplot2 theme  
  palette = c("#E7B800", "#2E9FDF"))
```


```
ggsurvplot(fit,data=dataRace,  
  pval = TRUE, conf.int = TRUE,  
  risk.table = TRUE, # Add risk table  
  risk.table.col = "strata", # Change risk table color by groups  
  linetype = "strata", # Change line type by groups  
  surv.median.line = "hv", # Specify median survival  
  ggtheme = theme_bw(), # Change ggplot2 theme  
  palette = c("#E7B800", "#2E9FDF"))
```

# ESTIMATION OF SURVIVAL DIFFERENCE BETWEEN WHITE AND BLACK OR AFRICAN AMERICAN





# Online databases



## Kaplan-Meier Plotter

Breast cancer

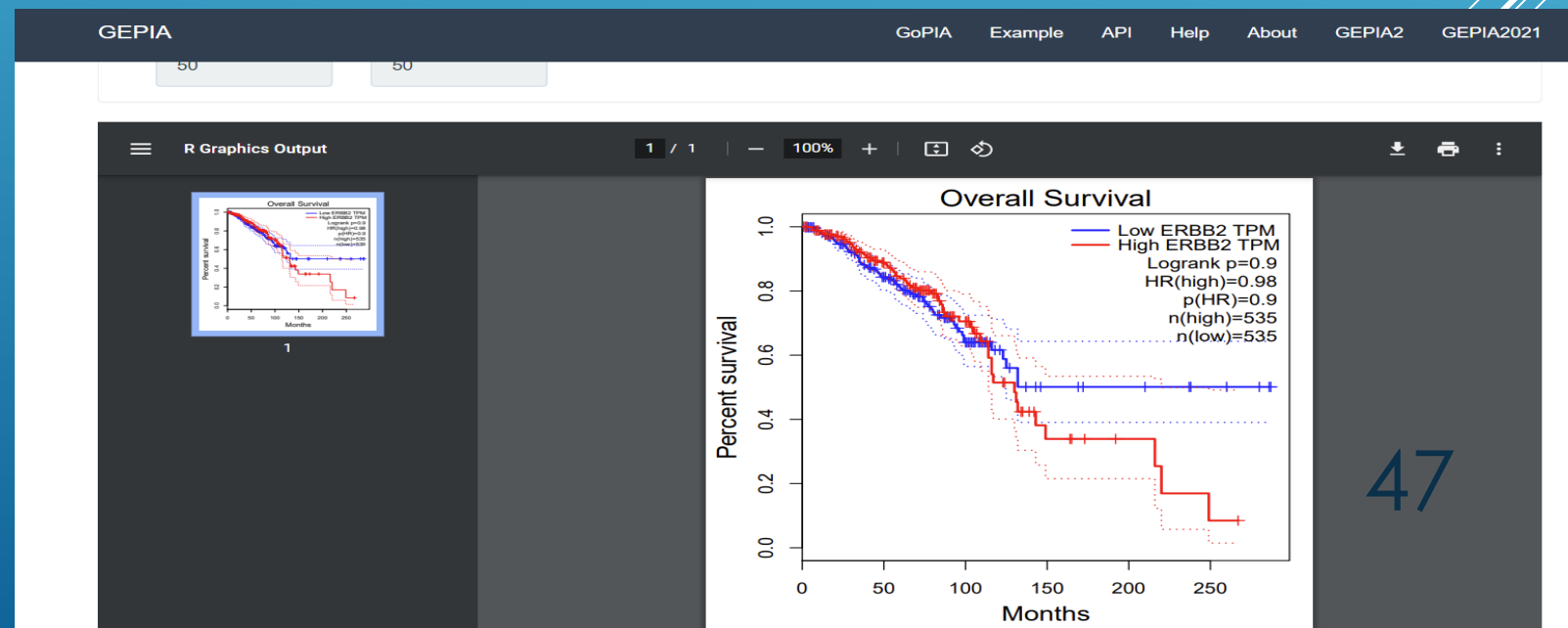
Breast cancer

**KM plotter** Home Upload Download Updates Contact

### What is the KM plotter?

The Kaplan Meier plotter is capable to assess the correlation between the expression of **all genes (mRNA, miRNA, protein)** and survival in **30k+ samples from 21 tumor types** including breast, ovarian, lung, & gastric cancer. Sources for the databases include GEO, EGA, and TCGA. Primary purpose of the tool is a meta-analysis based **discovery and validation of survival biomarkers** for cancer research.

<https://kmplot.com/analysis/>





**Any Questions**

# Thank You!

Email: [shivangi.agarwal800@gmail.com](mailto:shivangi.agarwal800@gmail.com)

Linkedin: <https://www.linkedin.com/in/shivangi-agarwal-4a6a3467/>