



A.D. 1308

unipg

UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Survival analysis I

Kaplan Meier and Log rank

PhD. Estevao Barcelos
v2025-02

Adapted from:

Babraham
Bioinformatics

Tutorial papers can be helpful for additional reading

- Clark, T., Bradburn, M., Love, S. et al. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 89, 232–238 (2003). <https://doi.org/10.1038/sj.bjc.6601118>
- Deo SV, Deo V, Sundaram V. Survival analysis-part 1. Indian J Thorac Cardiovasc Surg. 2020 Nov;36(6):668-672. doi: 10.1007/s12055-020-01049-1. Epub 2020 Oct 2. PMID: 33100633; PMCID: PMC7572944.

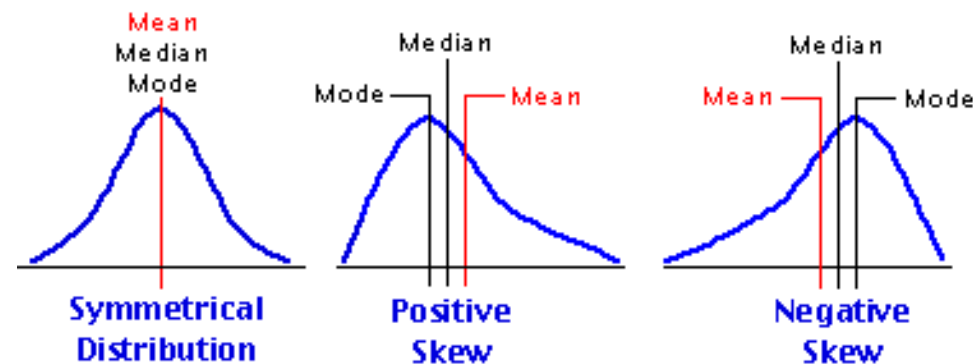
Outline - Survival analysis

- Time to **event data**.
- **Censoring**.
- **Survivor function – Kaplan-Meier estimator**.
- **Log-rank test**.

Time to event data: examples

- Time to death.
- Time to progression of cancer.
- Time to development of diabetes.
- Time to recovery from diarrhea.
- Time to event data typically collected in
 - cohort studies (time between study baseline and event of interest).
 - clinical trials (time between randomization and event of interest).
- Also known as **survival data**.

Features of time to event data

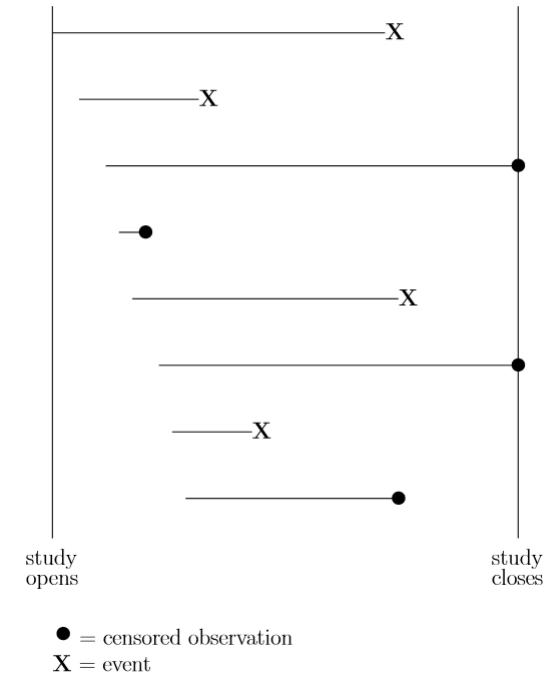


- Non-negative values.
- **Not normally distributed** (usually positively skewed).
- Event not usually observed for all individuals during the study.
- An observation is **censored** if individual does not experience event during the study.
- **Censoring time:** time from baseline/randomization until latest date at which individual is known to be still alive and event-free.

Censoring

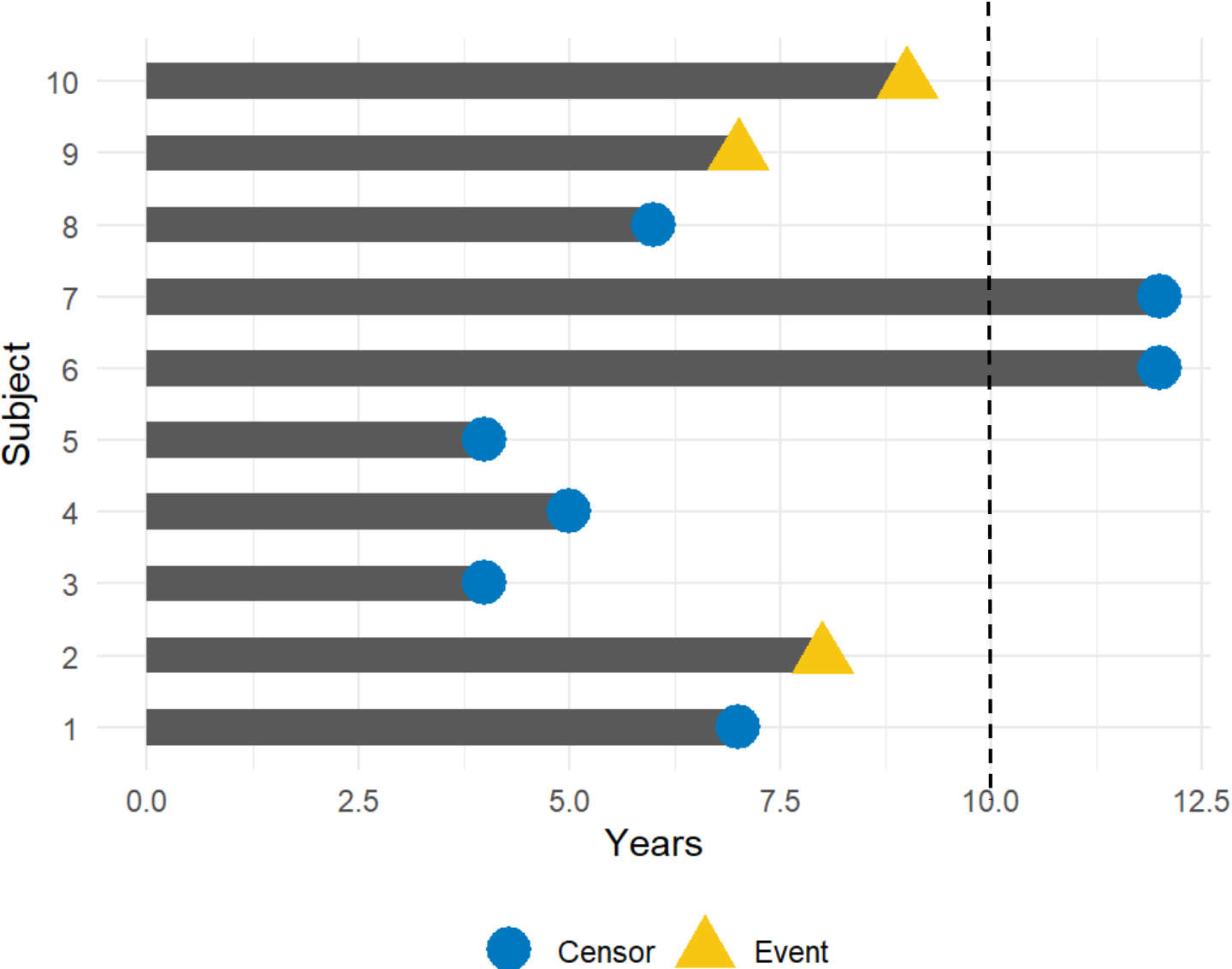
- Definition: Event of interest not observed for all individuals.
- **Fixed censoring:** event has not occurred when study has ended or data analysis is performed.
- **Loss to follow-up:** individual has been lost to follow-up (e.g. he/she no longer wishes to take part in study).

Illustration of survival data



- Survival analysis methods make use of information from censored observations.
- Assume censoring is **non-informative**, i.e. if an individual is censored, his/her subsequent risk of the event of interest is unaffected.

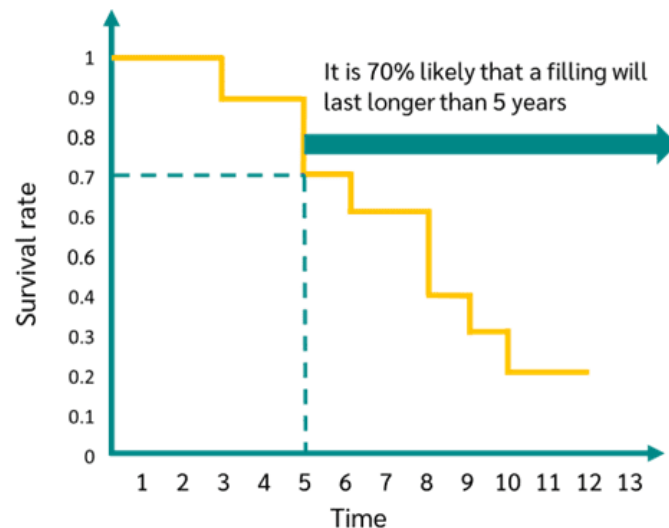
Censoring



Aims of survival analysis

- To estimate probability of not experiencing event of interest (not dying = “surviving”) over any given time period (e.g. 5 year survival rate).
- To compare overall survival experience between different groups of individuals (e.g. between groups in a randomized clinical trial).
- **Survivor function:** Probability of not experiencing event of interest (“surviving”) up to time t .

Example:



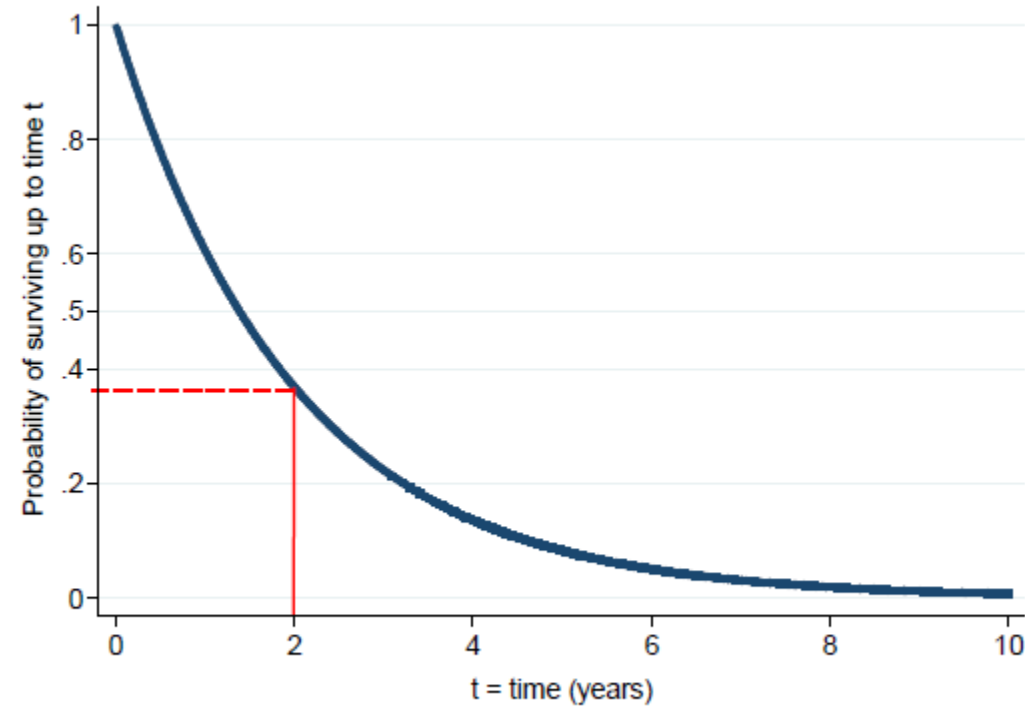
$$S(t) = P(T > t)$$

- T is the random variable representing the time until the event.
- $P(T > t)$ is the probability that the event has not yet occurred by time t .

Basically, $S(t)$ indicates how many people (or objects) are still "alive" or functional at time t .

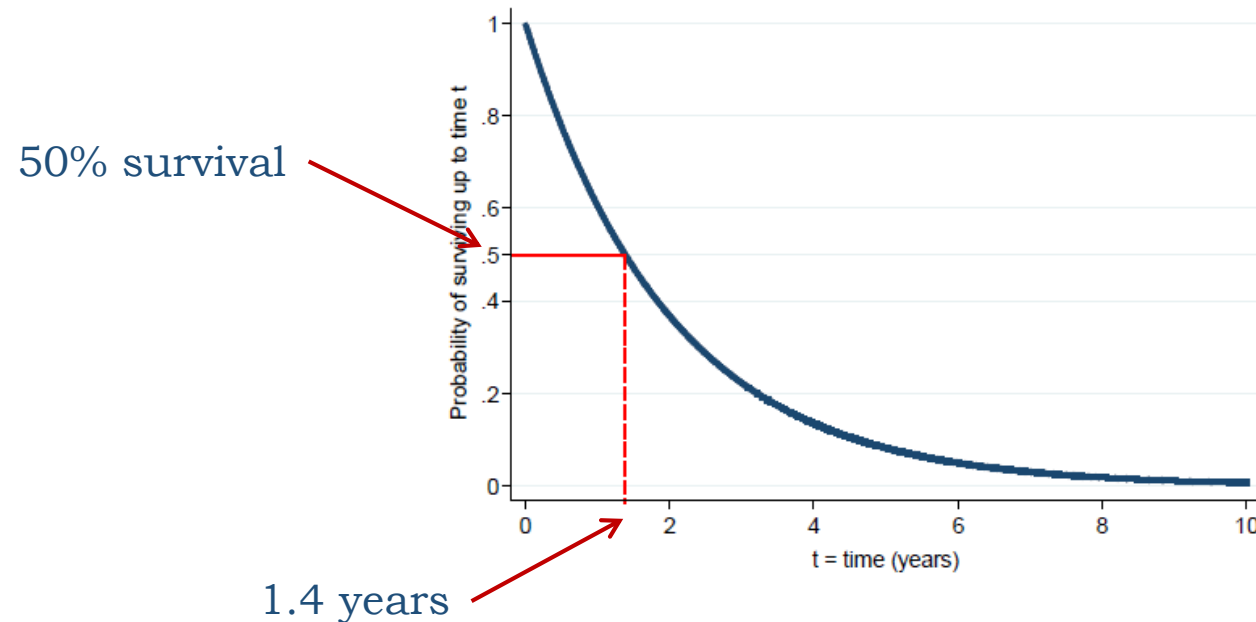
Estimating a survival rate

- Probability of surviving up to 2 years = 0.37.



Median survival time

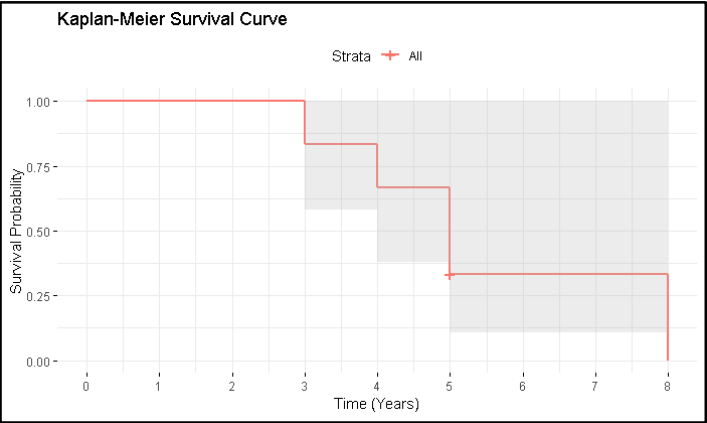
- It is the time (expressed in months or years) when half the patients are expected to be alive. It means that the chance of surviving beyond that time is 50%.
- Median survival time = 1.4 years, since the probability of surviving up to 1.4 years is 0.5.



Kaplan-Meier (KM) estimation of survivor function

- Is used to graphically represent the survival rate or survival function.
- **Non-parametric method.**

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$



Survival time (t)	Number at risk before the event (n)	Number of events at time t (d)	Fraction that has survived after event (1-d/n)	Survival probability
t	n	d	1-d/n	S(t)
3	5	1	1 - 1/5 = 0.80	1 x 0.8 = 0.8
4	4	1	1 - 1/4 = 0.75	0.8 x 0.75 = 0.6
5	3	2	1 - 2/3 = 0.33	0.6 x 0.33 = 0.2
8	1	1	1 - 1/1 = 0	0.2 x 0 = 0

In this fictitious table, we are assuming that we do not have censored data.

Example of time to event data

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

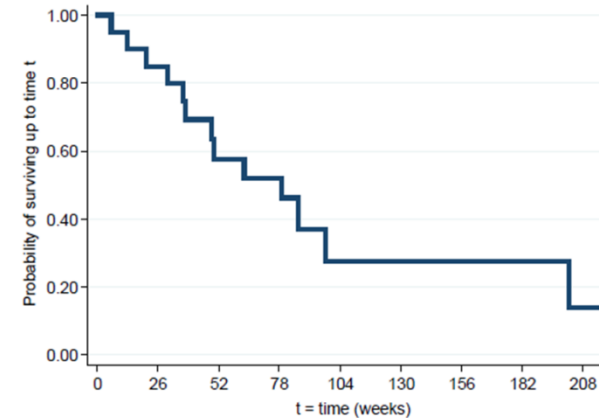
6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

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

Kaplan-Meier (KM) estimation of survivor function

First death

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



- **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 $(1-0.05)=0.95=19/20$.

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
0	20	0	0	1	1
6	20	1	0.05	0.95	$1 \times 0.95 = 0.95$

"Risk set" at time t

$1/20$

$19/20$

K-M estimation of survivor function

Second death

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

- **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: $\frac{1}{19}=0.053$ ^{$\frac{19}{20}$}
- Therefore probability of surviving beyond $t=13$ is $0.95 \times 0.947 = 0.90$. ^{$\frac{18}{19}$}
 - with $0.95=(1-(1/20))$ and $0.947=(1-(1/19))$

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
6	20	1	0.05	0.95	0.95
13	19	1	0.053	0.947	$0.95 \times 0.947 = 0.90$

$$\frac{1}{19}$$

$$1-(1/19)=18/19$$

K-M estimation of survivor function

Third and fourth death

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

- **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 **$0.90 \times (1-(1/18)) = 0.85$** .
- **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 **$0.85 \times (1-(1/17)) = 0.80$** .

From $t=13$: 0.95×0.947

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
13	19	1	$1/19 = 0.053$	0.947	0.90
21	18	1	$1/18 = 0.056$	0.944	0.85
30	17	1	$1/17 = 0.059$	0.941	0.80

K-M estimation of survivor function

Fifth and sixth death

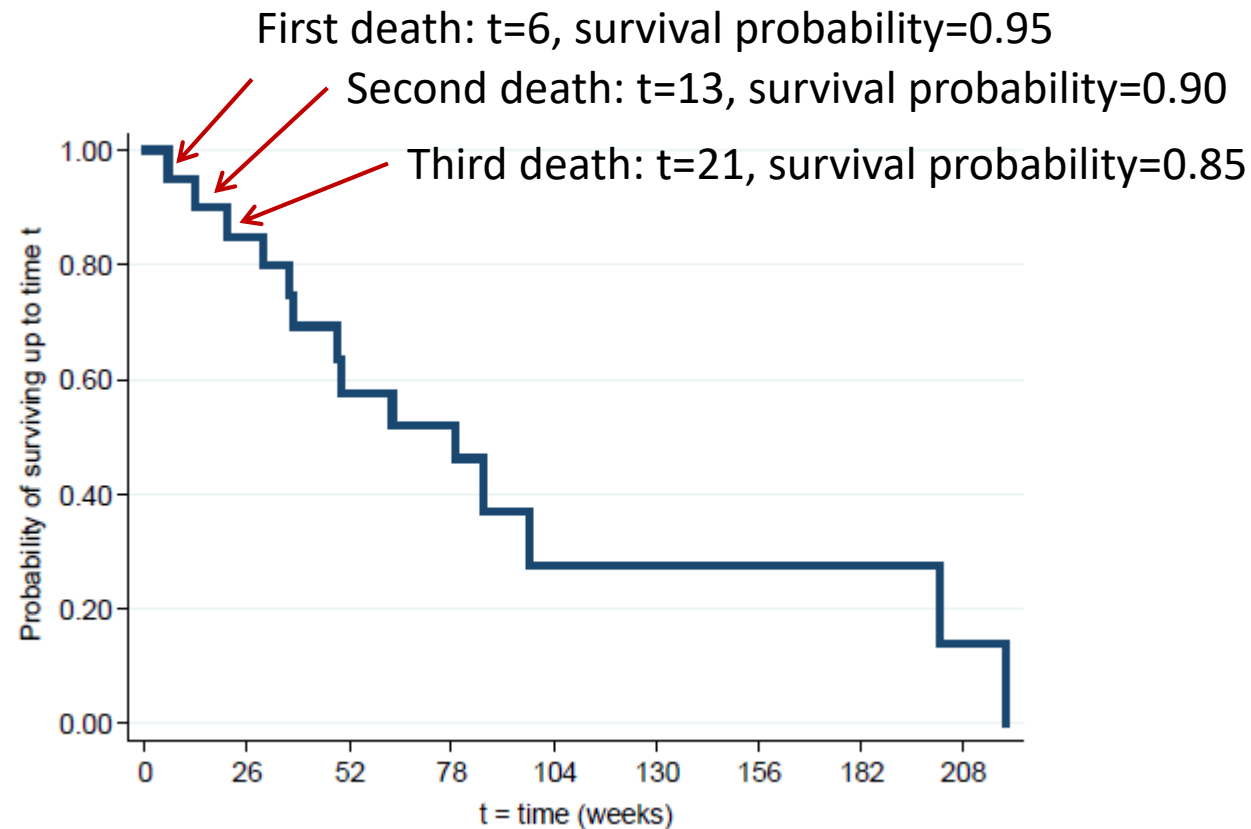
				31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219

- **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 **$0.80 \times (1 - (1/15)) = 0.747$.**

Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
30	17	1	0.059	0.941	0.80
31	16	0	0	1	$0.80 \times 1 = 0.80$
37	15	1	$1/15 = 0.067$	0.933	$0.80 \times 0.933 = 0.747$

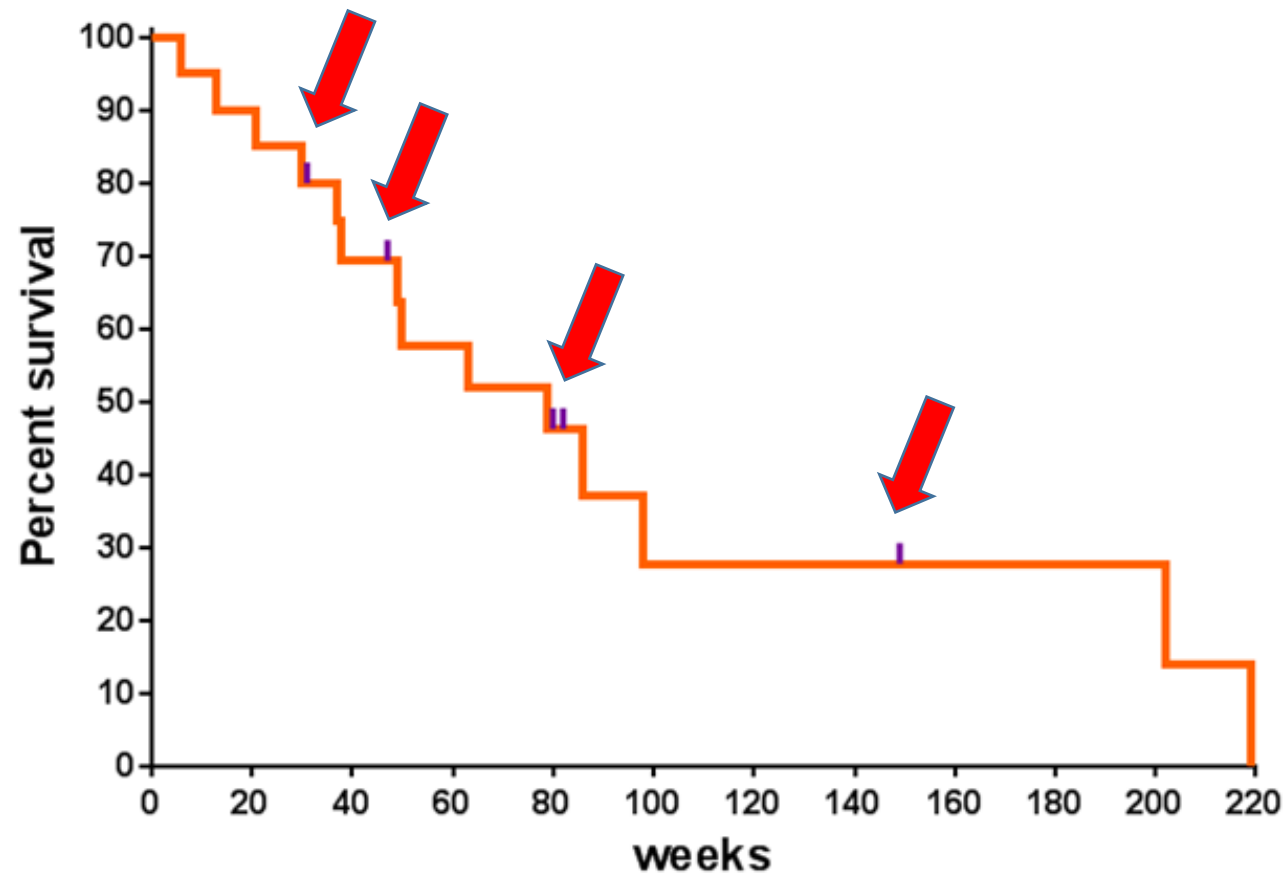
K-M plot of survivor function

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



K-M plot of survivor function

- Add ticks to indicate where censoring occurred.



Comparing 2 groups

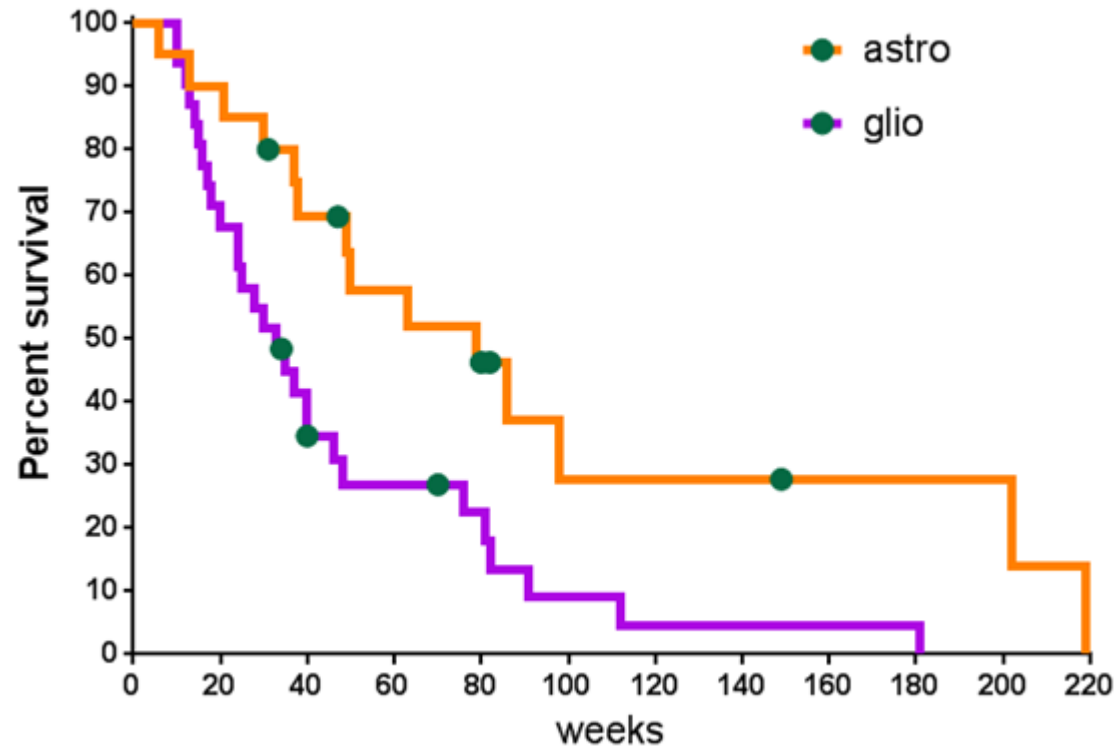
- Weeks to death or censoring (*) in **20 adults** with recurrent astrocytoma:

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									

K-M plot of survivor function by tumour type

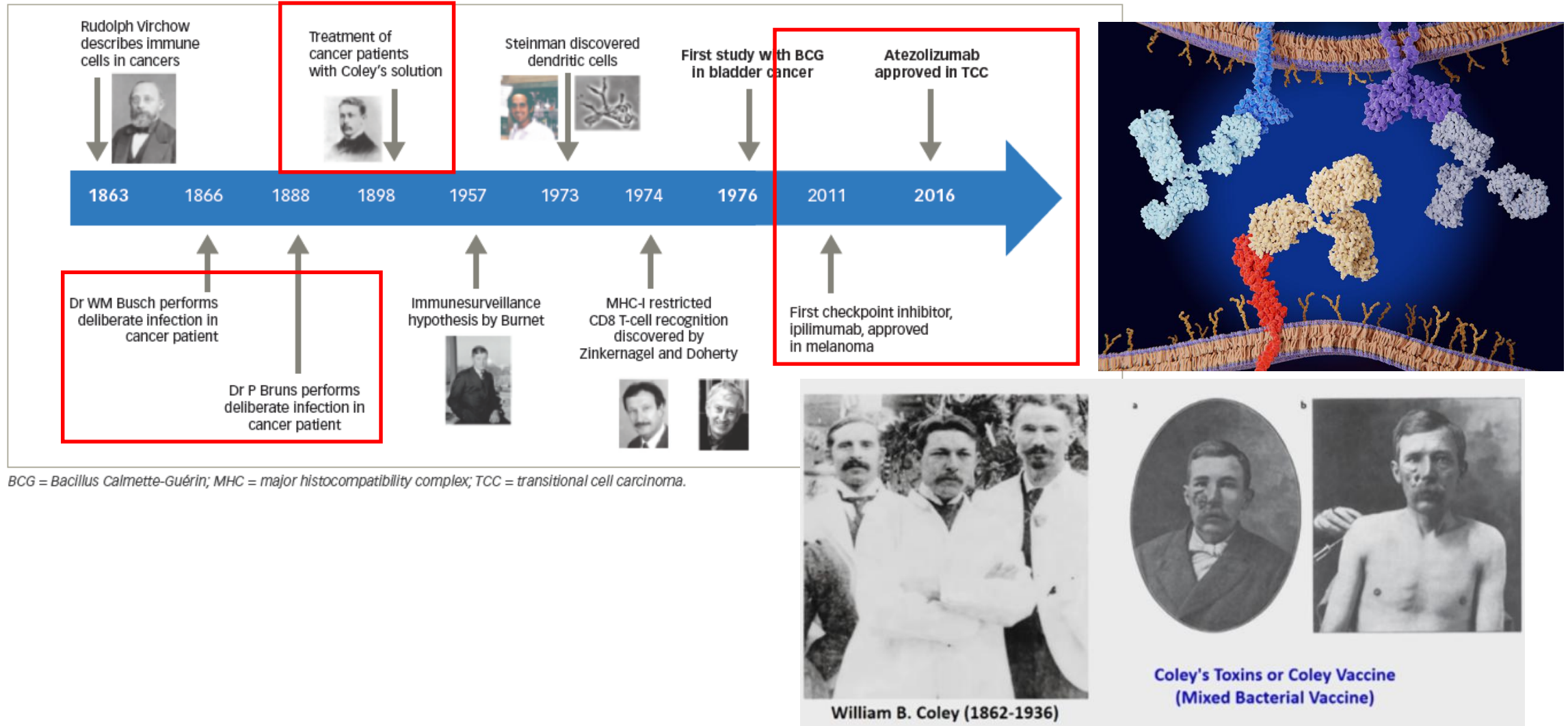


- Survival chances appear better in individuals with astrocytoma than with glioblastoma, but is the **difference between groups statistically significant?**

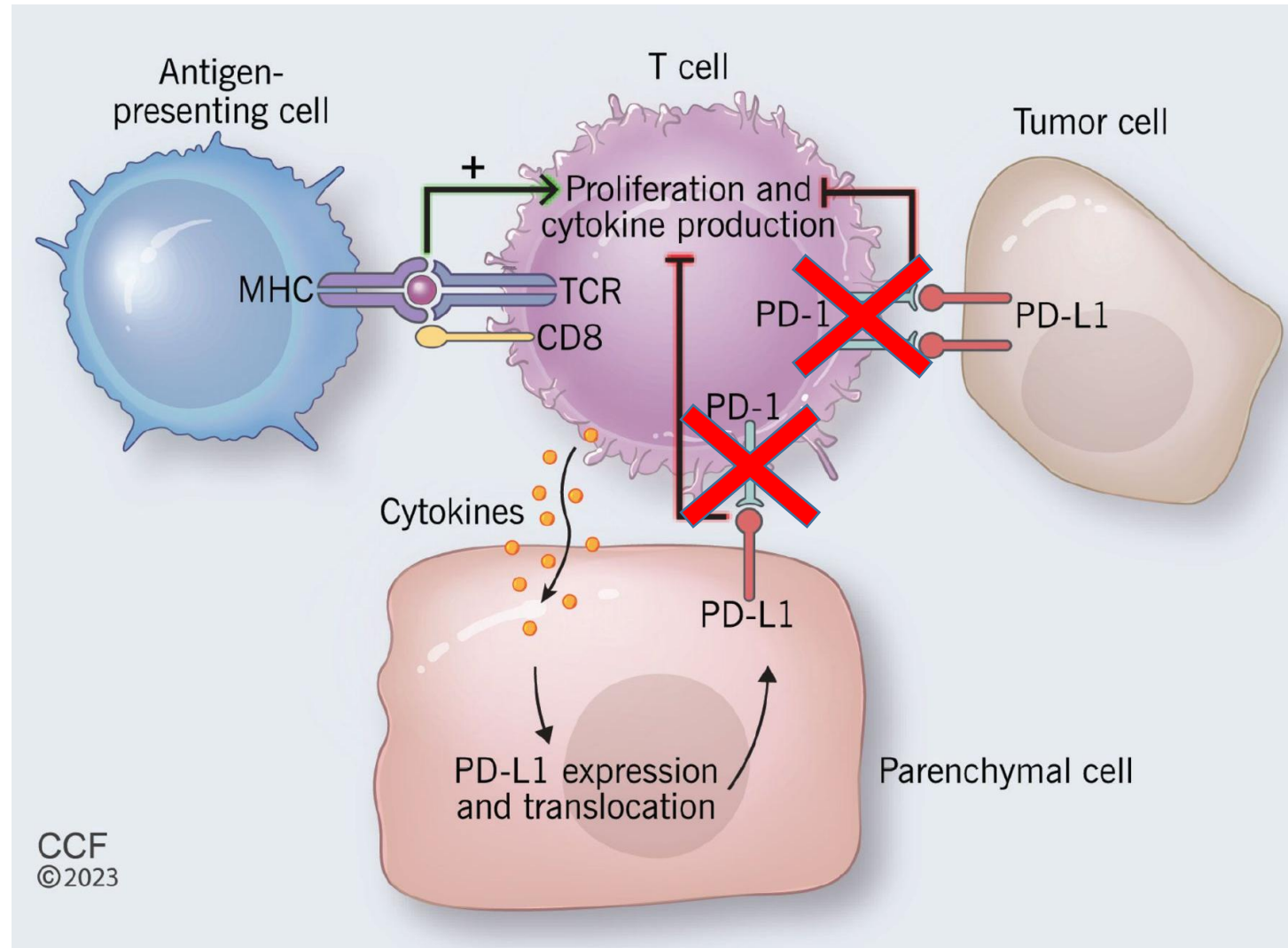
Comparing 2 samples

- Could compare **median survival time**, or **probability of surviving** up to any particular time.
- Better to use a test which compares survivor functions over whole follow-up period.
- **Log rank test: Non-parametric hypothesis 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.
- **Log rank test statistic:**
 - based on calculating expected number of events that would occur under null hypothesis at each event time, and comparing to observed number of events.
 - under null hypothesis has a Chi^2 distribution with 1 degree of freedom.

Cancer Immunotherapy: Past, Present, and Future



Cancer Immunotherapy: Past, Present, and Future



Chronology of use of ICI in esophageal adenocarcinoma

Nivolumab combined with chemotherapy for gastric, gastroesophageal junction and esophageal adenocarcinoma. Approval of the use of nivolumab + chemotherapy as first-line treatment for gastroesophageal adenocarcinomas with PD-L1 CPS ≥ 5 .



KEYNOTE-590

CheckMate 649

CheckMate 577

2020

2021-2022

2023 onwards

Chemotherapy, radiotherapy and surgery.

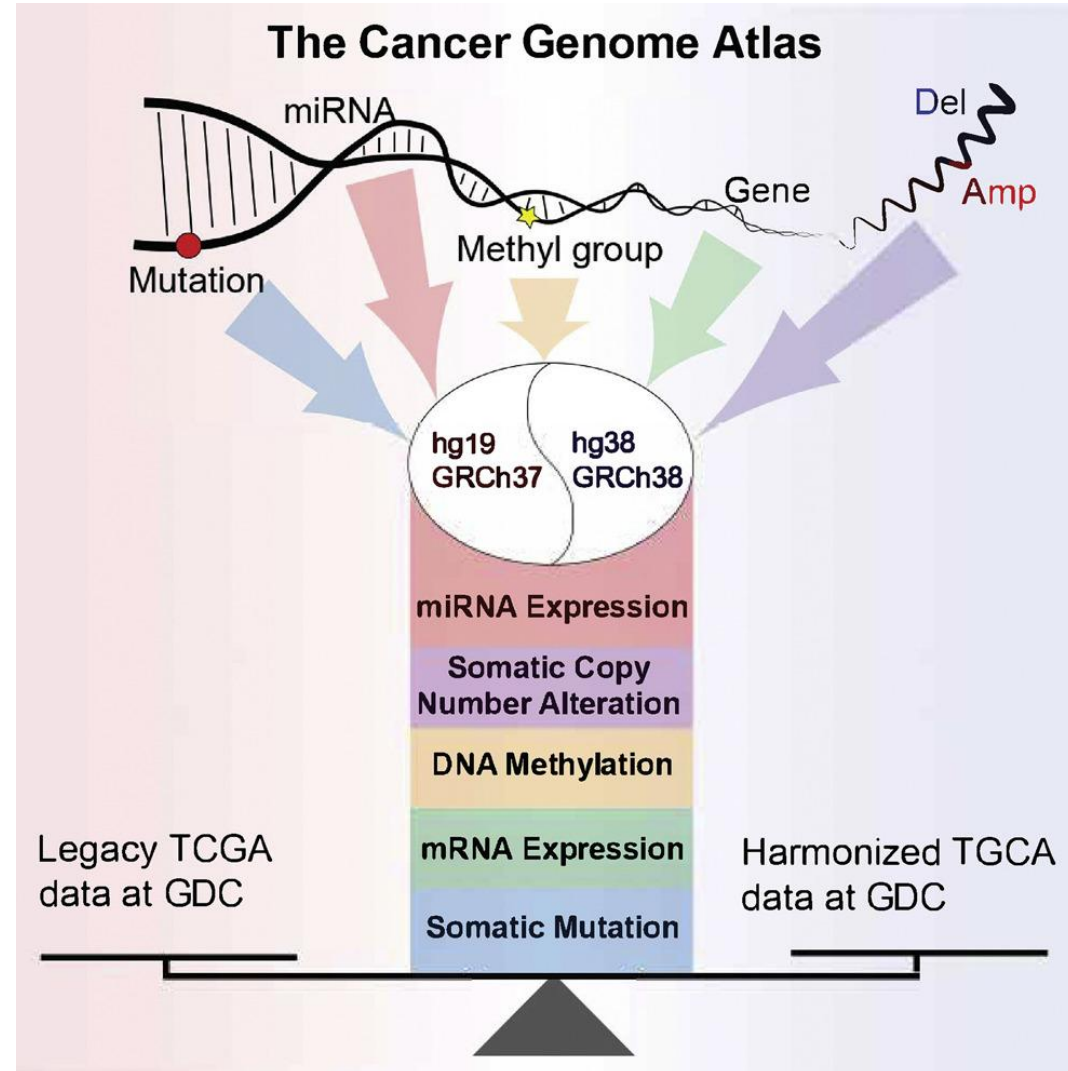
ICIs were being tested in clinical trials for advanced or metastatic gastroesophageal tumors.

FDA approved **pembrolizumab** in combination with chemotherapy as first-line treatment for PD-L1-positive advanced/metastatic esophageal cancer. CPS ≥ 10

Adjuvant **nivolumab** after chemoradiotherapy and surgery for esophageal cancer with incomplete pathological response.

$$\text{Combined Positive Score CPS} = \frac{\# \text{ PD-L1 positive cells (tumor cells, lymphocytes, macrophages)}}{\# \text{ viable tumor cells}} \cdot 100$$

TCGA ESCA - Esophageal Carcinoma





<https://xena.ucsc.edu/>

[Overview](#)[Analysis](#)[Tutorials](#)[What's New](#)[Cite Us](#)[Subscribe](#)

UCSC Xena

See the bigger picture

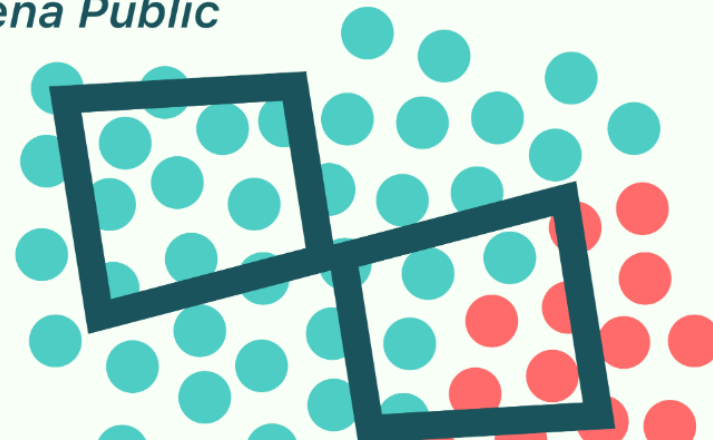
An online exploration tool for public and private,
multi-omic and clinical/phenotype data

Launch Xena

Tutorials and walkthroughs

Don't know where to start? Jump in with one of

Xena Public



237 Cohorts, 2253 Datasets

[10x visium Human Ovarian Cancer \(7 datasets\)](#)
[10x visium Human Ovarian Cancer enhanced resolution \(3 datasets\)](#)
[10x visium Mouse Brain Coronal \(6 datasets\)](#)
[10x visium Mouse Sagittal Anterior1 \(6 datasets\)](#)
[10x visium V1_Breast_Cancer_Block_A_Section_1 \(4 datasets\)](#)
[Acute lymphoblastic leukemia \(Mullighan 2008\) \(3 datasets\)](#)
[Breast Cancer \(Caldas 2007\) \(3 datasets\)](#)
[Breast Cancer \(Chin 2006\) \(3 datasets\)](#)
[Breast Cancer \(Haverty 2008\) \(2 datasets\)](#)
[Breast Cancer \(Hess 2006\) \(2 datasets\)](#)
[Breast Cancer \(Miller 2005\) \(2 datasets\)](#)
[Breast Cancer \(vantVeer 2002\) \(2 datasets\)](#)
[Breast Cancer \(Vijver 2002\) \(2 datasets\)](#)
[Breast Cancer \(Yau 2010\) \(2 datasets\)](#)
[Breast Cancer Cell Lines \(Heiser 2012\) \(4 datasets\)](#)
[Breast Cancer Cell Lines \(Neve 2006\) \(2 datasets\)](#)
[Cancer Cell Line Encyclopedia \(Breast\) \(4 datasets\)](#)
[Cancer Cell Line Encyclopedia \(CCLE\) \(9 datasets\)](#)
[Connectivity Map \(2 datasets\)](#)
[cSCC \(19 datasets\)](#)
[cSCC scRNA-seq \(7 datasets\)](#)
[cSCC_scRNAseq_visium \(6 datasets\)](#)
[follicular_remodeling \(4 datasets\)](#)
[GBM \(Parsons 2008\) \(2 datasets\)](#)
[GDC APOLLO-LUAD \(8 datasets\)](#)
[GDC BEATAML1.0-COHORT \(7 datasets\)](#)
[GDC CDDP_EAGLE-1 \(9 datasets\)](#)
[GDC CGCI-BLGSP \(10 datasets\)](#)
[GDC CGCI-HTMCP-CC \(11 datasets\)](#)
[GDC CGCI-HTMCP-DLBCL \(9 datasets\)](#)
[GDC CGCI-HTMCP-LC \(9 datasets\)](#)
[GDC CMI-ASC \(6 datasets\)](#)
[GDC CMI-MBC \(6 datasets\)](#)
[GDC CMI-MPC \(6 datasets\)](#)
[GDC CPTAC-2 \(7 datasets\)](#)
[GDC CPTAC-3 \(12 datasets\)](#)
[GDC CTSP-DI BCI 1 \(6 datasets\)](#)

Active Data Hubs

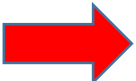
- ☒ [My computer hub](#)
- ☒ [UCSC Public Hub](#)
- ☒ [TCGA Hub](#)
- ☒ [Pan-Cancer Atlas Hub](#)
- ☒ [ICGC Hub](#)
- ☒ [PCAWG Hub](#)
- ☒ [UCSC Toil RNA-seq Recompute](#)
- ☐ [Treehouse Hub](#)
- ☒ [GDC Hub](#)
- ☒ [ATAC-seq Hub](#)
- ☒ [Kids First Hub](#)
- ☒ <https://previewsinglecell.xenahubs.net>
- ☒ [jupyter notebook](#)

[TCGA Bile Duct Cancer \(CHOL\) \(17 datasets\)](#)
[TCGA Bladder Cancer \(BLCA\) \(21 datasets\)](#)
[TCGA Breast Cancer \(BRCA\) \(24 datasets\)](#)
[TCGA Cervical Cancer \(CESC\) \(18 datasets\)](#)
[TCGA Colon and Rectal Cancer \(COADREAD\) \(15 datasets\)](#)
[TCGA Colon Cancer \(COAD\) \(25 datasets\)](#)
[TCGA Endometrioid Cancer \(UCEC\) \(26 datasets\)](#)
[TCGA Esophageal Cancer \(ESCA\) \(19 datasets\)](#)
[TCGA Formalin Fixed Paraffin-Embedded Pilot Phase II \(FPPP\) \(2 datasets\)](#)
[TCGA Glioblastoma \(GBM\) \(25 datasets\)](#)
[TCGA Head and Neck Cancer \(HNSC\) \(21 datasets\)](#)
[TCGA Kidney Chromophobe \(KICH\) \(18 datasets\)](#)
[TCGA Kidney Clear Cell Carcinoma \(KIRC\) \(23 datasets\)](#)
[TCGA Kidney Papillary Cell Carcinoma \(KIRP\) \(21 datasets\)](#)
[TCGA Large B-cell Lymphoma \(DLBC\) \(17 datasets\)](#)

[VISUALIZE](#)

cohort: TCGA Esophageal Cancer (ESCA)

copy number (gene-level)



gene expression RNAseq

[IlluminaHiSeq* \(n=196\)](#) TCGA Hub

The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates, as in $\log_2(x+1)$ transformed RSEM normalized count. Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details). Reference to method description from University of North Carolina TCGA genome characterization center: DCC description. In order to more easily view the differential expression between samples, we set the default view to center each gene or exon to zero by independently subtracting the mean of each gene or exon on the fly. Users can view the original non-normalized values by adjusting visualization settings.

[IlluminaHiSeq \(n=198\)](#) TCGA Hub

The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the British Columbia Cancer Agency TCGA genome characterization center. Level 3 data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription estimates, as in RPKM values (Reads Per Kilobase of exon model per Million mapped reads). Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details). In order to more easily view the differential expression between samples, we set the default view to center each gene or exon to zero by independently subtracting the mean of each gene or exon on the fly. Users can view the original non-normalized values by adjusting visualization settings.

[IlluminaHiSeq pancan normalized \(n=196\)](#) TCGA Hub

TCGA esophageal carcinoma (ESCA) gene expression by RNAseq, mean-normalized (per gene) across all TCGA cohorts. Values in this dataset are generated at UCSC by combining "gene expression RNAseq" values of all TCGA cohorts, values are then mean-centered per gene, then extracting the converted data only belongs to the this cohort. For comparing data within this cohort, we recommend to use the "gene expression RNAseq" dataset. For questions regarding the gene expression of this particular cohort in relation to other types tumors, you can use the pancan normalized version of the "gene expression RNAseq" data. For comparing with data outside TCGA, we recommend using the percentile version if the non-TCGA data is normalized by percentile ranking. For more information, please see our Data FAQ: [here](#).

[IlluminaHiSeq percentile \(n=196\)](#) TCGA Hub

For each sample, we rank genes RSEM values between 0% to 100%. This dataset is gene expression estimation in percentile rank, which higher value representing higher expression. The dataset can be used to compare this RNAseq data with other cohorts when the other data is processed in the same way (i.e. percentile ranking). Genes are mapped onto the human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details). Reference to method description from University of North Carolina TCGA genome characterization center: DCC description. For comparing data within this cohort, we recommend to use the "gene expression RNAseq" dataset. For questions regarding the gene expression of this particular cohort in relation to other types tumors, you can use the pancan normalized version of the "gene expression RNAseq" data. For comparing with data outside TCGA, we recommend using the percentile version if the non-TCGA data is normalized by percentile ranking. For more information, please see our Data FAQ: [here](#). In order to more easily view the differential expression between samples, we set the default view to center each gene or exon to zero by independently subtracting the mean of each gene or exon on the fly. Users can view the original non-normalized values by adjusting visualization settings.

miRNA mature strand expression RNAseq

[IlluminaHiSeq \(n=195\)](#) TCGA Hub

pathway activity

[Paradigm IPLs \(n=181\)](#) TCGA Hub

The PARADIGM algorithm integrates pathway, expression and copy number data to infer activation of pathway features within a superimposed pathway (SuperPathway) network structure. The SuperPathway structure comprises 1500 constituent pathways from three pathway databases, NCI-PID, BioCarta and Reactome (last updated 05/2013), containing 19K pathway features, representing 7369 genes, 9354 complexes, 2092 families, 82 RNAs, 15 miRNAs and 592 abstract processes. This dataset is the PARADIGM integrated pathway levels (IPLs) of ~19K pathway features of whitelisted Pan-Cancer 33 samples from this cohort, computed using the platform corrected RNA-seq and GISTIC thresholded gene level copy number data.

[z score of 1387 constituent PARADIGM pathways \(n=181\)](#) TCGA Hub

The PARADIGM algorithm integrates pathway, expression and copy number data to infer activation of pathway features within a superimposed pathway (SuperPathway) network structure. The SuperPathway structure comprises 1387 constituent pathways from three pathway databases, NCI-PID, BioCarta and Reactome (last updated 05/2013), containing 19K pathway features, representing 7369 genes, 9354 complexes, 2092 families, 82 RNAs, 15 miRNAs and 592 abstract processes. This dataset is ssGSEA scores for 1387 constituent pathways, z transformed

phenotype

[Curated survival data \(n=204\)](#) TCGA Hub

Curated survival data from the Pan-cancer Atlas paper titled "An Integrated TCGA Pan-Cancer Clinical Data Resource (TCGA-CCR) to drive high quality survival outcome analytics". The paper highlights four types of carefully curated survival endpoints, and recommends the use of the endpoints of OS, PFI, DFI, and DSS for each TCGA cancer type. OS: overall survival PFI: progression-free interval DSS: disease-specific survival DFI: disease-free interval

[Phenotypes \(n=204\)](#) TCGA Hub

protein expression DDPA

dataset: gene expression RNAseq - IlluminaHiSeq

hub: <https://tcga.xenahubs.net>

TCGA esophageal carcinoma (ESCA) gene expression by RNAseq (polyA+ IlluminaHiSeq)

The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform by the University of North Carolina TCGA genome characterization dataset shows the gene-level transcription estimates, as in $\log_2(x+1)$ transformed RSEM normalized count. Genes are mapped onto the human genome coordinates using UCSC Xena method description from University of North Carolina TCGA genome characterization center: [DCC description](#)

In order to more easily view the differential expression between samples, we set the default view to center each gene or exon to zero by independently subtracting the mean of each gene adjusting visualization settings.

cohort	TCGA Esophageal Cancer (ESCA)
dataset ID	TCGA.ESCA.sampleMap/HiSeqV2
download	https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.ESCA.sampleMap%2FHiSeqV2.gz ; Full metadata
samples	196
version	2017-10-13
type of data	gene expression RNAseq
unit	$\log_2(\text{norm_count}+1)$
platform	IlluminaHiSeq_RNASeqV2
ID/Gene Mapping	https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/probeMap%2Fhugo_gencode_good_hg19_V24lift37_probemap ; Full metadata
author	University of North Carolina TCGA genome characterization center
raw data	https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/esca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/
wrangling	Level_3 data (file names: *.rsem.genes.normalized_results) are downloaded from TCGA DCC, $\log_2(x+1)$ transformed, and processed at UCSC into Xena repos
input data format	ROWS (identifiers) x COLUMNS (samples) (i.e. genomicMatrix)

20,531 identifiers X 196 samples [All Identifiers](#) [All Samples](#)

	TCGA-L5-A4OF-01	TCGA-LN-A49R-01	TCGA-Z6-A9VB-01	TCGA-L5-A8NF-01	TCGA-2H-A9GG-01	TCGA-IG-A3QL-01	TCGA-1
? 100130426	0.2799	0.000	0.000	0.4466	0.000	0.000	
? 100133144	5.251	5.241	5.620	5.085	4.565	4.238	
? 100134869	5.083	5.976	5.599	5.020	3.569	4.172	
? 10357	2.475	2.474	5.078	5.507	4.924	2.241	
? 10431	8.939	9.765	8.821	9.748	9.757	10.24	
? 136542	0.000	0.000	0.000	0.000	0.000	0.000	

sample	_PATIENT	OS	OS.time	DSS	DSS.time	DFI	DFI.time	PFI	PFI.time
TCGA-2H-A9GF-01	TCGA-2H-A9GF	1	784	1	784		1	172	
TCGA-2H-A9GG-01	TCGA-2H-A9GG	1	610	1	610		1	504	
TCGA-2H-A9GH-01	TCGA-2H-A9GH	1	951	1	951		1	800	
TCGA-2H-A9GI-01	TCGA-2H-A9GI	1	435	1	435		1	400	
TCGA-2H-A9GJ-01	TCGA-2H-A9GJ	1	1781	1	1781		1	1261	
TCGA-2H-A9GK-01	TCGA-2H-A9GK	1	232	1	232		1	113	
TCGA-2H-A9GL-01	TCGA-2H-A9GL	1	180	1	180		1	109	
TCGA-2H-A9GM-01	TCGA-2H-A9GM	1	424	1	424		1	367	
TCGA-2H-A9GN-01	TCGA-2H-A9GN	1	272	1	272		1	242	
TCGA-2H-A9GO-01	TCGA-2H-A9GO	1	494	1	494		1	362	
TCGA-2H-A9GQ-01	TCGA-2H-A9GQ	1	128	1	128		1	112	
TCGA-2H-A9GR-01	TCGA-2H-A9GR	1	987	1	987		1	987	
TCGA-IC-A6RE-01	TCGA-IC-A6RE	0	234	0	234	0	234	0	234
TCGA-IC-A6RE-11	TCGA-IC-A6RE	0	234	0	234	0	234	0	234
TCGA-IC-A6RF-01	TCGA-IC-A6RF	0	477	0	477	1	293	1	293
TCGA-IC-A6RF-11	TCGA-IC-A6RF	0	477	0	477	1	293	1	293
TCGA-IG-A3I8-01	TCGA-IG-A3I8	0	1012	0	1012	0	1012	0	1012
TCGA-IG-A3I8-11	TCGA-IG-A3I8	0	1012	0	1012	0	1012	0	1012
TCGA-IG-A3QL-01	TCGA-IG-A3QL	0	1071	0	1071	0	1071	0	1071
TCGA-IG-A3Y9-01	TCGA-IG-A3Y9	1	26	0	26	0	26	0	26
TCGA-IG-A3YA-01	TCGA-IG-A3YA	0	632	0	632	0	632	0	632
TCGA-IG-A3YB-01	TCGA-IG-A3YB	0	80	0	80	0	80	0	80
TCGA-IG-A3YC-01	TCGA-IG-A3YC	0	612	0	612		1	542	
TCGA-IG-A4P3-01	TCGA-IG-A4P3	1	567	1	567		1	567	
TCGA-IG-A4Q5-01	TCGA-IG-A4Q5	1	118	0	118	0	118	0	118
TCGA-IG-A4QT-01	TCGA-IG-A4QT	1	283	0	283	0	283	0	283
TCGA-IG-A50L-01	TCGA-IG-A50L	0	16	0	16		0	16	
TCGA-IG-A51D-01	TCGA-IG-A51D	0	518	0	518		0	518	
	TCGA-2H-A9GN-01	null	null	null	1.000		272.0	1.000	272.0
	TCGA-2H-A9GO-01	null	null	null	1.000		494.0	1.000	494.0

dataset: phenotype - Phenotypes

hub: <https://tcga.xenahubs.net>

cohort [TCGA Esophageal Cancer \(ESCA\)](#)

dataset ID TCGA.ESCA.sampleMap/ESCA_clinicalMatrix

download https://tcga-xena-hub.s3.us-east-1.amazonaws.com/download/TCGA.ESCA.sampleMap%2FESCA_clinicalMatrix; Full metadata

samples 204

version 2019-12-06

type of data phenotype

raw data https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/esca/bcr/







input data format ROWs (samples) x COLUMNs (identifiers) (i.e. clinicalMatrix)

204 samples X 121 identifiers [All Identifiers](#) [All Samples](#)

	CDE_ID_3226963	_GENOMIC_ID_TCGA_ESCA_PDMRNAseq	_GENOMIC_ID_TCGA_ESCA_PDMRNAseqCNV	_GENOMIC_ID_TCGA_ESCA_RPPA	_GENOMIC_ID_TCGA_ESCA_exp_HiSeq	_GENOMIC_ID_TCGA_ESCA_exp_HiSeqV2	C
TCGA-2H-A9GF-01	MSI-L	TCGA-2H-A9GF-01	TCGA-2H-A9GF-01	7F800F35-6E6A-4CEA-AAC7-09745EE5A79D	TCGA-2H-A9GF-01A-11R-A37I-31	2666431c-515b-4088-a448-baf7c52106d8	
TCGA-2H-A9GG-01	MSI-L	TCGA-2H-A9GG-01	null	null	TCGA-2H-A9GG-01A-11R-A37I-31	9fcd3933-2651-4c64-8c77-3dac4d8ea595	
TCGA-2H-A9GH-01	MSS	TCGA-2H-A9GH-01	TCGA-2H-A9GH-01	null	TCGA-2H-A9GH-01A-11R-A37I-31	775bb1a6-0463-4e78-bc3f-596e3a5eb38f	
TCGA-2H-A9GI-01	MSS	TCGA-2H-A9GI-01	TCGA-2H-A9GI-01	1565C181-0C35-45E9-B549-D5909F672417	TCGA-2H-A9GI-01A-11R-A37I-31	3e9a405f-cb27-4773-b501-57a2eaf833d4	
TCGA-2H-A9GJ-01	MSS	TCGA-2H-A9GJ-01	TCGA-2H-A9GJ-01	null	TCGA-2H-A9GJ-01A-11R-A37I-31	67bda296-5e65-45a9-962e-a99f51a12e14	
TCGA-2H-A9GK-01	MSS	TCGA-2H-A9GK-01	TCGA-2H-A9GK-01	null	TCGA-2H-A9GK-01A-11R-A37I-31	fd1ad4dc-ff32-48df-bed7-45fd742418e9	
TCGA-2H-A9GL-01	MSS	TCGA-2H-A9GL-01	TCGA-2H-A9GL-01	1481BCF2-2B38-4D7C-BB4D-7DCFD7B272F1	TCGA-2H-A9GL-01A-12R-A37I-31	9b3adc67-cc9c-48db-96d4-2314368986e2	
TCGA-2H-A9GM-01	MSS	TCGA-2H-A9GM-01	TCGA-2H-A9GM-01	null	TCGA-2H-A9GM-01A-11R-A37I-31	7b4fe474-b28b-466b-9985-8c34ef90fccf	
TCGA-2H-A9GN-01	MSS	TCGA-2H-A9GN-01	TCGA-2H-A9GN-01	null	TCGA-2H-A9GN-01A-11R-A37I-31	3b88034b-950d-4fd1-ae98-2cd37342bbf4	
TCGA-2H-A9GO-01	MSS	TCGA-2H-A9GO-01	TCGA-2H-A9GO-01	null	TCGA-2H-A9GO-01A-11R-A37I-31	b4ab4edc-6163-4211-80c2-12d8bbbd8a2a	

Files to be used in the Analysis



 survival_ESCA_survival		11/04/2025 09:49	Documento di testo	11 KB
 TCGA.ESCA.sampleMap_ESCA_clinicalMatrix		11/04/2025 09:49	File SAMPLEMAP_...	222 KB
 TCGA.ESCA.sampleMap_HiSeqV2		11/04/2025 09:48	WinRAR archive	10.663 KB



Posit

<https://posit.co> › [download](#) › [rst...](#) · [Traduzir esta página](#) ⋮



RStudio Desktop

1: Install R. **RStudio requires R 3.6.0+**. Choose a version of R that matches your computer's operating system. R is not a Posit product ...

1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.

DOWNLOAD AND INSTALL R

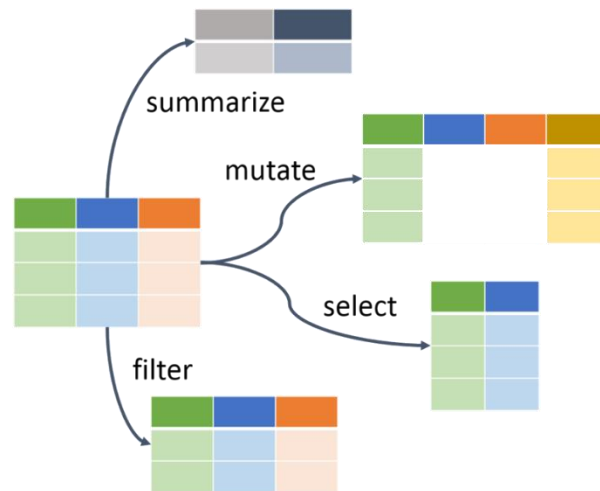
2: Install RStudio

DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS

Size: 265.28 MB | [SHA-256: BB369743](#) | Version: 2024.12.1+563 |
Released: 2025-02-13

Libraries

```
#-----  
# Survival Analysis - TCGA Data (RSEM log2(x+1) normalized values)  
#-----  
  
# Install required libraries  
install.packages("dplyr")  
install.packages("survival")  
install.packages("survminer")  
  
# Load required libraries  
library(dplyr)  
library(survival)  
library(survminer)
```



survival package

Surv(time, event)

Survfit()

coxph()

survminer package

ggsurvplot()

ggforest()

Count Matrix and Metadata

```
# Part 1. Count Matrix and Metadata.
# -----

# Load Count Matrix.
count_matrix <- read.delim(file.choose(), header = TRUE, row.names = 1, sep = "\t")
colnames(count_matrix) <- gsub("\\.", "-", colnames(count_matrix))

# Load Phenotype Data.
clinical_data <- read.delim(file.choose(), header = TRUE, row.names = 1)
survival_data <- read.delim(file.choose(), header = TRUE, row.names = 1)
survival_data <- survival_data[, -ncol(survival_data)]

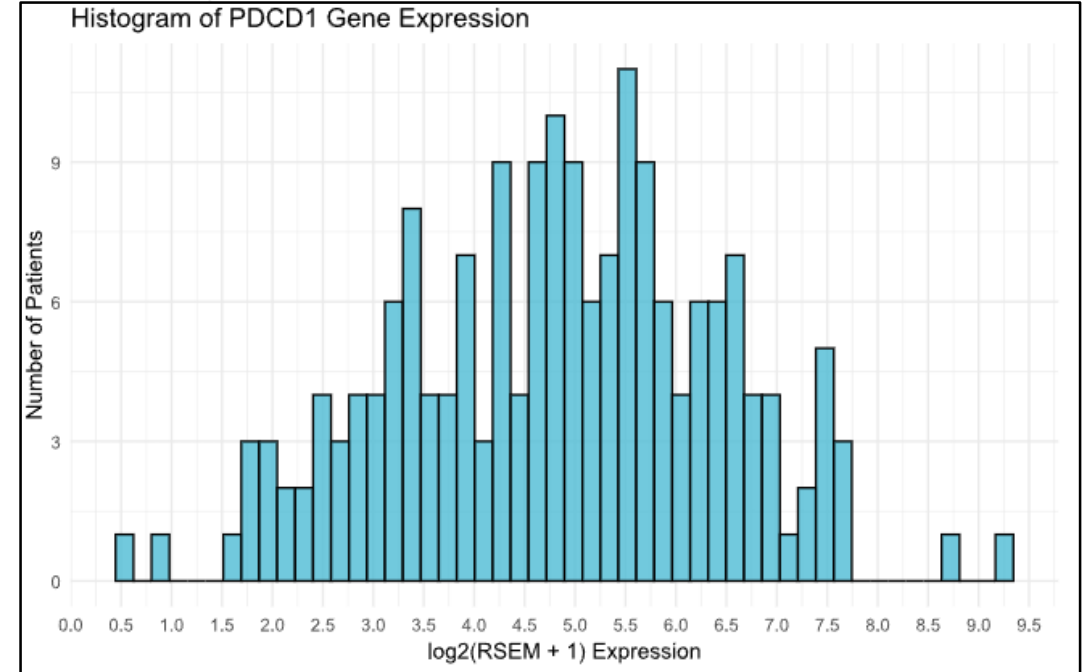
# Merge clinical and survival data
merged_data <- merge(clinical_data, survival_data, by = "row.names", all = TRUE)
rownames(merged_data) <- merged_data$Row.names
merged_data$Row.names <- NULL

# Filter for Primary Tumor and/or Metastatic samples
merged_data <- merged_data %>% filter(sample_type %in% "Primary Tumor")

# Identify common patients
common_patients <- intersect(colnames(count_matrix), rownames(merged_data))
merged_data <- merged_data %>% filter(rownames(merged_data) %in% common_patients)
count_matrix <- count_matrix[, common_patients]
```

Gene expression Analysis

```
# Part 2. Gene signature Analysis.  
# -----  
  
# Extract gene expression values from count_matrix  
gene_sig_exp <- count_matrix["PDCD1", ]  
  
# Ensure gene_exp is a numeric vector and sample IDs match between merged_data and count_matrix  
gene_sig_exp <- as.numeric(gene_sig_exp)  
  
# Match and correctly order the gene expression values with merged_data's sample IDs  
merged_data$gene_sig_exp <- gene_sig_exp[match(rownames(merged_data), colnames(count_matrix))]  
  
# Verify that gene_exp is correctly added  
print(summary(merged_data$gene_sig_exp))  
head(merged_data)
```



Cut-off Median-based

```
# Part 3. Cut-off Median
# -----

# Calculate the median expression of the gene signature
median_gene_signature_exp <- median(merged_data$gene_sig_exp, na.rm = TRUE)

# Split patients into low and high expression groups based on the median
merged_data$expression_group <- ifelse(merged_data$gene_sig_exp >
                                       median_gene_signature_exp, "High", "Low")

# Check the distribution of groups
table(merged_data$expression_group)

# Verify the first few rows
head(merged_data)
```

Plot Overall Survival

```
# Part 4. Overall Survival Analysis
# -----

# Fit survival model
surv_object <- Surv(time = merged_data$OS.time, event = merged_data$OS)
fit <- survfit(surv_object ~ expression_group, data = merged_data)

# Plot Kaplan-Meier survival curves
plot1 <- ggsurvplot(fit, data = merged_data, pval = TRUE, risk.table = TRUE,
                    title = expression(OS ~ "- PD1 expression"),
                    xlab = "Time (days)", ylab = "Overall Survival Probability",
                    legend.labs = c("High Expression", "Low Expression"),
                    palette = c("#d1495b", "#2e4057"),
                    break.time.by = 365)

# Rotate x-axis labels diagonally
plot1$plot <- plot1$plot + theme(axis.text.x = element_text(angle = 45, hjust = 1))
plot1$table <- plot1$table + theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Show survival plot
plot1
```

Plot Disease Specific Survival

```
# Part 5. Disease Specific Survival Analysis
# -----

# Fit survival model
surv_object <- Surv(time = merged_data$DSS.time, event = merged_data$DSS)
fit <- survfit(surv_object ~ expression_group, data = merged_data)

# Plot Kaplan-Meier survival curves
plot2 <- ggsurvplot(fit, data = merged_data, pval = TRUE, risk.table = TRUE,
                    title = expression(DSS ~ "- PD1 expression"),
                    xlab = "Time (days)", ylab = "Disease Specific Survival Probability",
                    legend.labs = c("High Expression", "Low Expression"),
                    palette = c("#d1495b", "#2e4057"),
                    break.time.by = 365)

# Rotate x-axis labels diagonally
plot2$plot <- plot2$plot + theme(axis.text.x = element_text(angle = 45, hjust = 1))
plot2$table <- plot2$table + theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Show survival plot
plot2

# End of script.
#####
```

Plot Final Result

OS - PD1 expression

