Survival Analysis of cancer patients using ML techniques
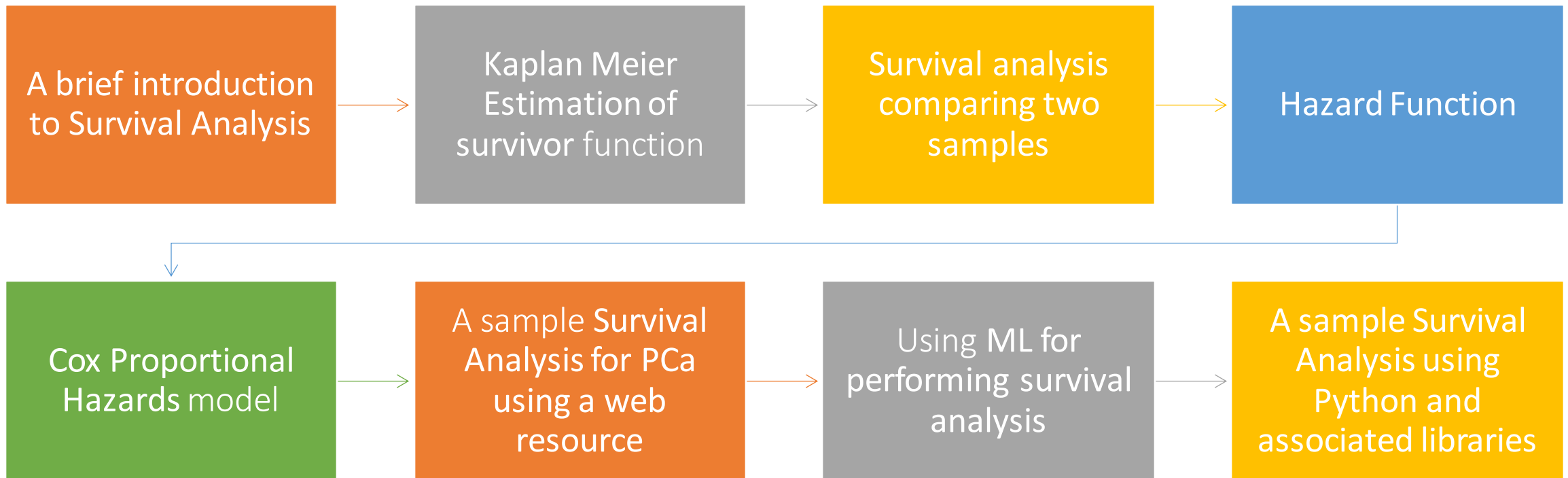
Shakuntala Baichoo, PhD

Assoc. Prof, Dept. Digital Technologies, **Uni. Of Mauritius**

Visiting Scientist, Dana-Farber Cancer Institute, Harvard University

ICTS MLHD (24-Jul to 4-Aug 2023), ICTS-TIFR, Bengaluru (India)

Agenda

A brief introduction to Survival Analysis	Kaplan Meier Estimation of survivor function	Survival analysis comparing two samples	Hazard Function

Cox Proportional Hazards model	A sample Survival Analysis for PCa using a web resource	Using ML for performing survival analysis	A sample Survival Analysis using Python and associated libraries

What is Survival Analysis?

Survival analysis, aka time-to-event analysis, is a statistical data analysis technique, designed to analyse the amount of time it takes for an event to occur, over an observation period.

It is utilized in biology, medicine, engineering, marketing, social sciences or behavioral sciences.(PdM)).

The units of time in survival analysis range from days, weeks, months, years and even decades from the beginning of follow up till an event occurs or until observation ceases (censorship).

Some events of interest

Death of a person by any disease

Attrition of employees from organization

An incident of a natural calamity (earthquake or flood).

Failure of machine tools

What is Survival analysis?

- Survival refers to probabilities.
  - Probability of an event occurrence after a certain time is **survival probability**.

- For instance, when we say the survival probability of a breast cancer patient surviving 7 years after undergoing mastectomy is 0.56.
  - This means that the patient will survive at least 7 years after a mastectomy and there is a probability 0.56 that she will keep surviving after 7 years.

- Similarly, there could be a probability 0.8 that she will survive after 3 years since mastectomy.
  - **Reason is clear**. Within 3 years, chances of disease recurrence and/or metastasis is lower and thus survival probability will be higher.

- A sample table for survival probabilities

Patient	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Survival (Months)	140.5	84.6	163.7	164.9	41.4	7.8	164.3	22.4	99.5	47.5
Probability of Survival	0.5	0.4	0.6	0.7	0.2	0.1	0.4	0.3	0.5	0.3

Survival Analysis of cancer patients

In many cancer studies, the main outcome under assessment is the time to an event of interest. Time survived from complete remission to relapse/progression of disease. Time from diagnosis to death.

At the end of follow-up some of the individuals may not have the event of interest, and thus their true time to event is unknown.

Survival data are rarely Normally distributed but are skewed and comprise typically of many early events and relatively few late ones.

It is these features of the data that make the special methods called survival analysis necessary.

Censoring

Most survival analyses must consider a very important analytical problem called **censoring**.

It is caused by not observing some subjects for the full time till failure(or event).

Say we have records for patients dying from breast cancer, but in some situations, it may not be possible to mark the exact time of death.

Problem occurs when the patient dies in between, after the end of or before the study, and hence censoring occurs.

There can be three primary reasons for this:

Patient does not have the event (death) before the study ends.
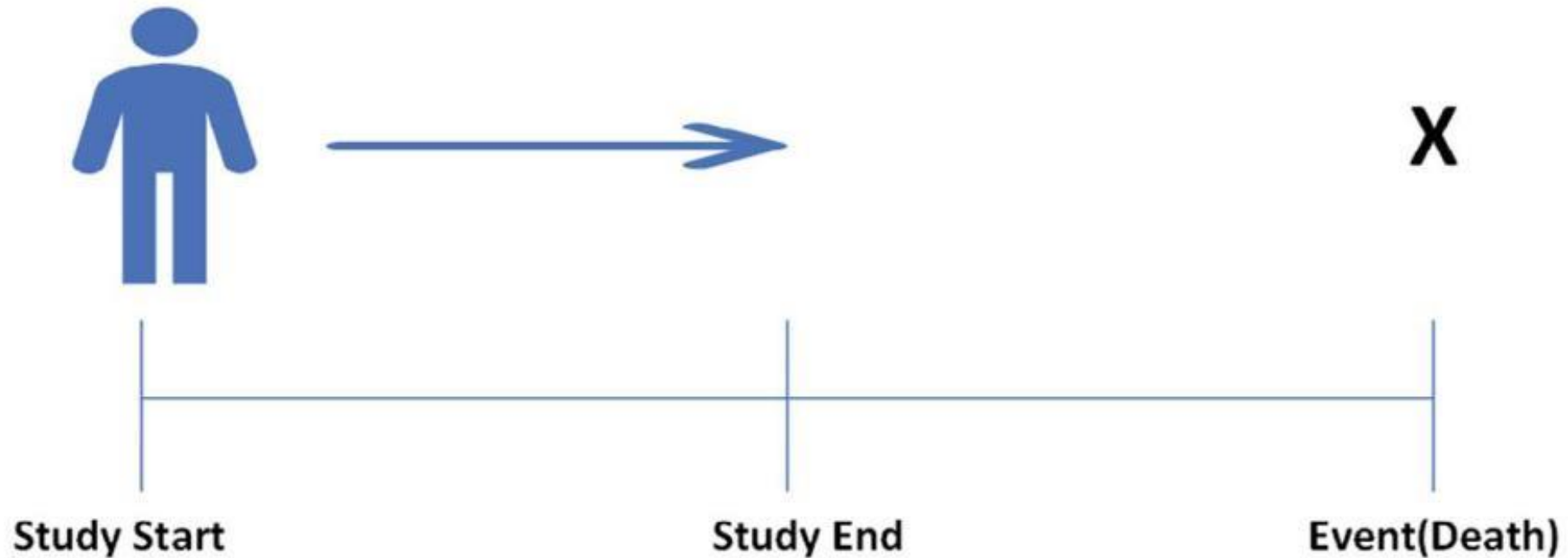
Patient left follow-up during the study period.

Patient died in the study period.

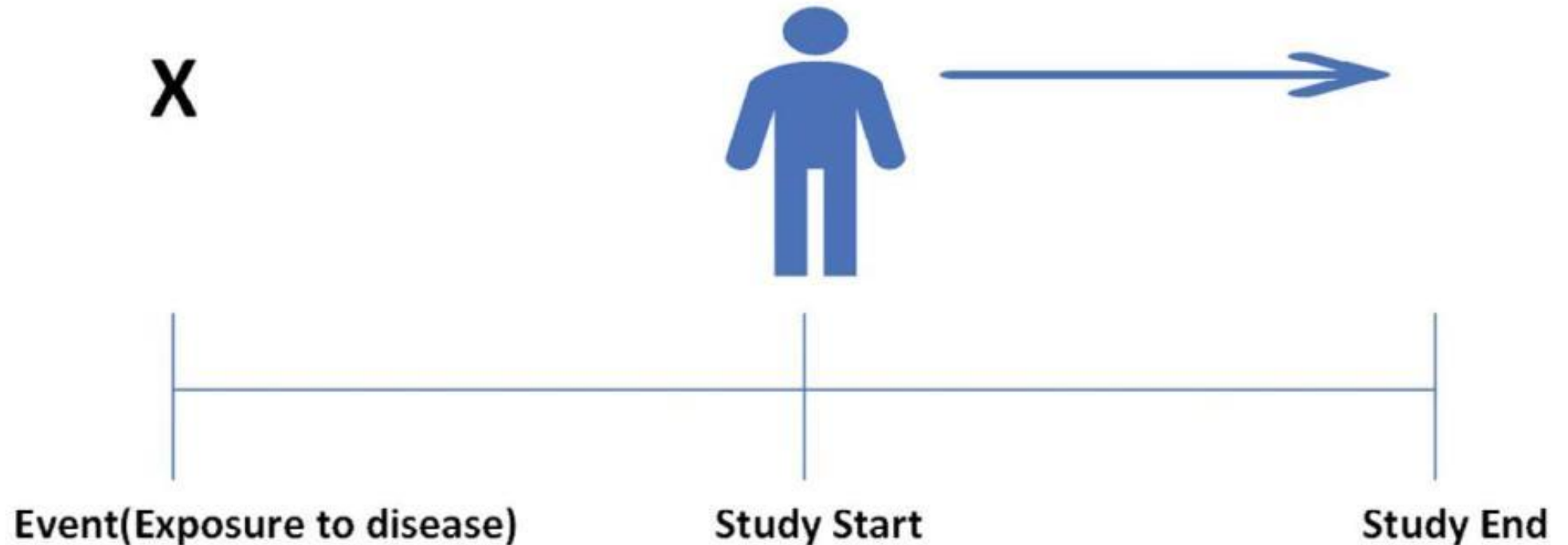# Right Censoring

Study ends but no event is observed.

Hence, true survival time is greater or equal to the observed time. It is the most popular censoring mechanism.

It is the most popular censoring mechanism.

| Study Start | Study End | Event(Death) |

Left Censoring

Event has already occurred before the start of the study.

True survival time is less than or equal to the observed time.

This occurs when someone comes for a follow-up and first exposure is noted as an event.

X

| Event(Exposure to disease) | Study Start | Study End |

Interval Censoring

Event occurred somewhere between time t1 and t2.

This scheme is applicable to both death and exposure to disease events.

It is a combination of both left and right censoring with one limit as infinite.

Study Start t_1 Event (Death) t_2 Study End
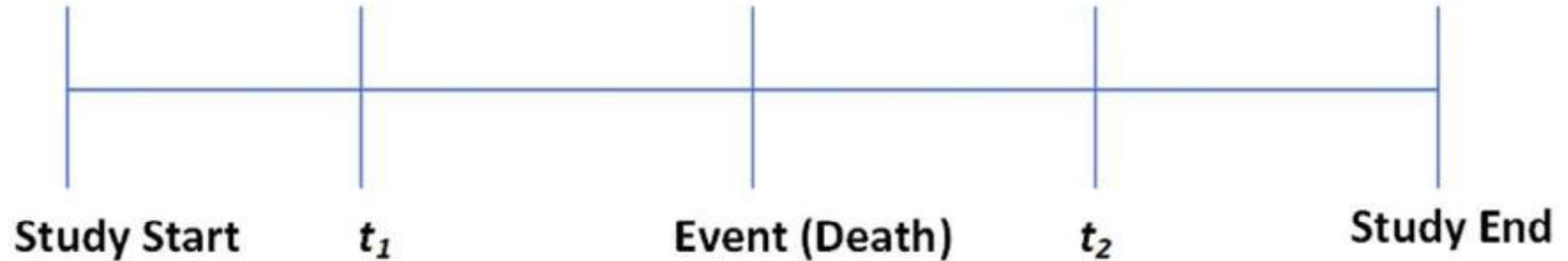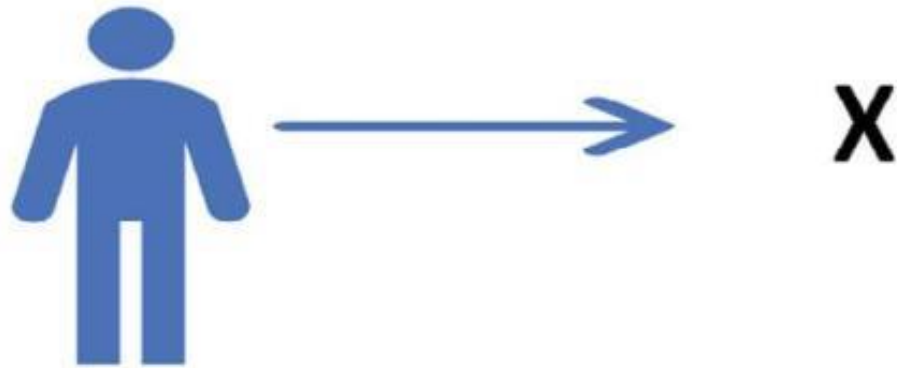
How are survival data described?

- Described and modelled in terms of two related probabilities, namely *survival* and *hazard*.

- The survival probability (which is also called the **survivor function**) $S(t)$ is the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified future time t.
  - It is fundamental to a survival analysis because survival probabilities for different values of t provide crucial summary information from time to event data.

- These values describe directly the survival experience of a study cohort.

Survivor function

survivor function returns the probability of an event occurring after time t.

Mathematically,

$$S(t) = P(T > t), 0 < t < \infty$$

Alternatively, it can be said that S(t) gives us the probability of a subject surviving after time t.

S(t) is nothing but a probability distribution over time.

S(t) is the probability of survival to time t (that is, the probability of no event before time t).

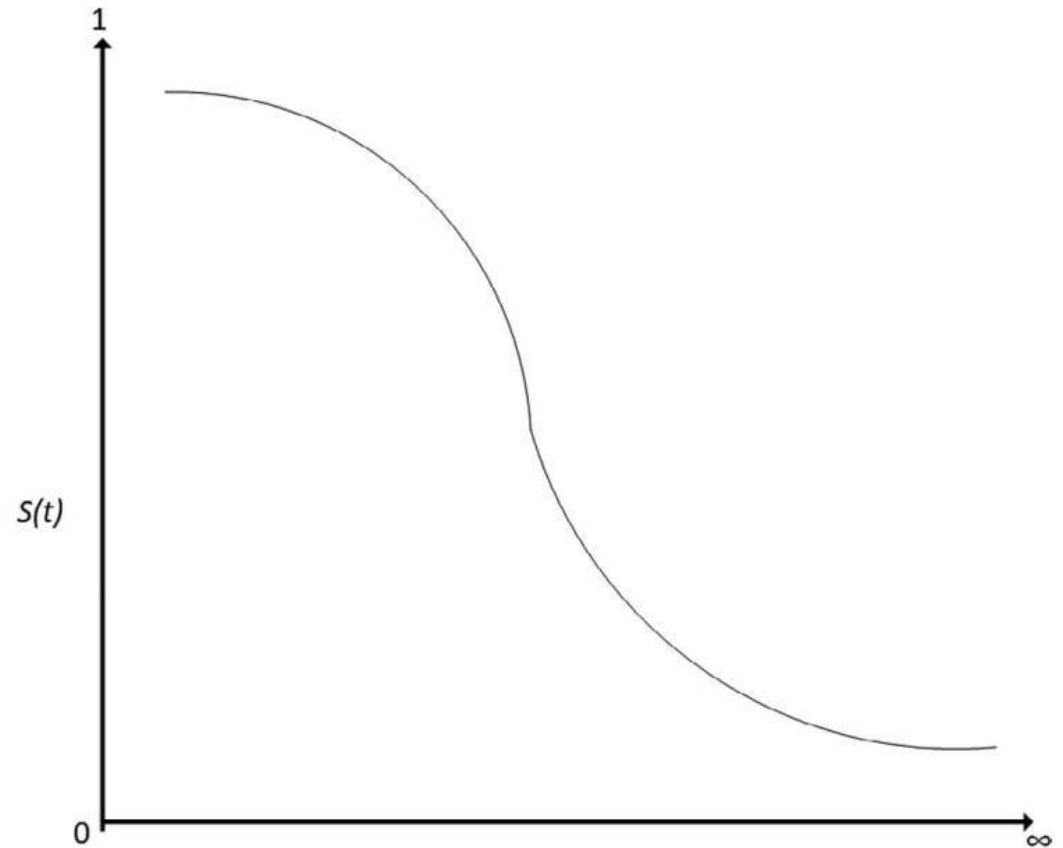Theoretically, t ranges from 0 to infinity, and of course, S(t) will have values from 0 to 1.

At the starting time, t_0, the survival probability is $S(t_0) = 1$ and normally, $S(\infty) = 0$, meaning that an event eventually occurs.

Ideally, survivor is represented by a decreasing smooth curve which begins at S(t) = 1 at t = 0.

Survivor function

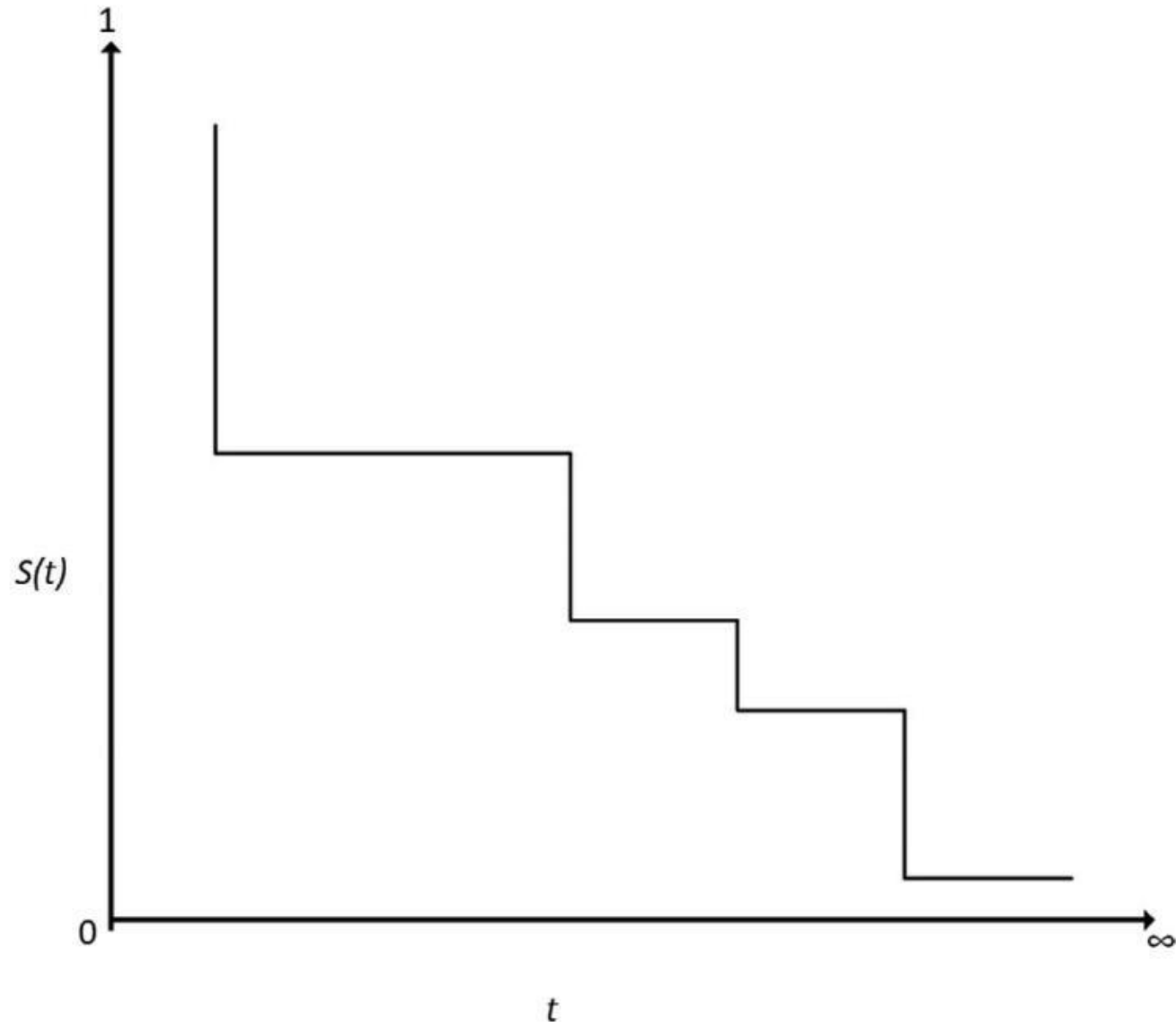- Theoretically, survivor looks as follows:

Survivor function

In practice, survival curves generated from real datasets look more broken and stepwise.

Survivor function

One important feature of the survivor function is that it is monotically decreasing,

- $S(t_1) < S(t_2)$ where $t_1 > t_2$

Survival chances from an event tend to decrease over time.

- e.g., chances of surviving of a cancer patient (from event 'Death') decreases as the time passes.

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 randomised clinical trial or patients harbouring mutations in a specific set of genes).

Kaplan Meier Estimation of Survivor Function

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

- It is a non-parametric technique of estimating and plotting the survival probability as a function of time.

- Let's consider a sample dataset from:
    - Bland JM, Altman DG. The logrank test. BMJ. 2004 May 1;328(7447):1073. doi: 10.1136/bmj.328.7447.1073. PMID: 15117797; PMCID: PMC403858

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

Kaplan Meier Estimation of Survivor Function – <u>first death</u>

6	13	21	30	31*	37	38	47*	49	50
63	79	80*	82*	82*	86	98	149	202	219
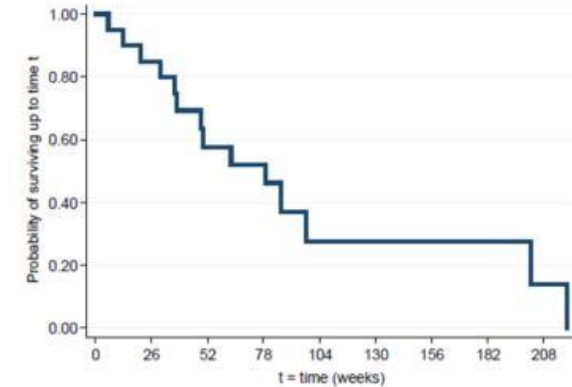
- **20** individuals in study at t=0.
- First death at t=6 weeks.
- No individuals censored before t=6.
- Probability of death for each individual: **1/20=0.05**
- Therefore probability of surviving beyond t=6 is **(1-0.05)=0.95=19/20**.

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 x 0.95 = 0.95

"Risk set" at time t   1/20   19/20

***

Kaplan Meier 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: **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))**

19/20   18/19

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 x 0.947 = 0.90

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

***

Kaplan Meier Estimation of survivor – 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 x (1-(1/18)) =0.85**.

From t=13: 0.95*0.947

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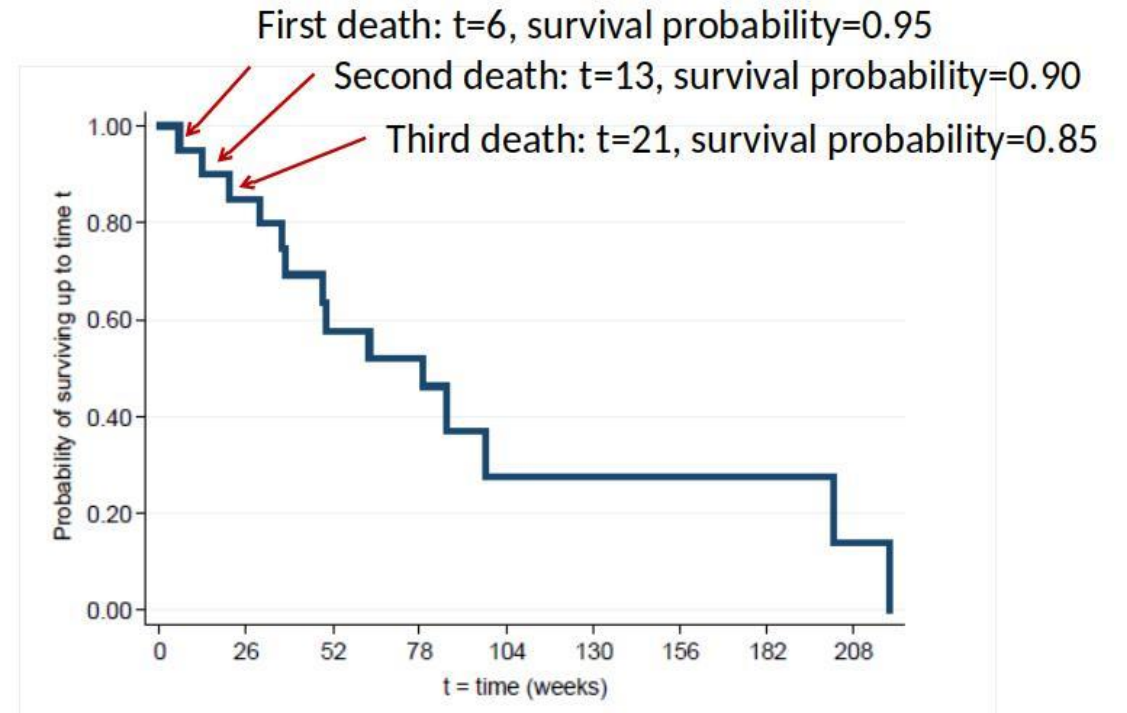- Probability of surviving beyond t=30 is **0.85 x (1-(1/17)) =0.80**.

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

***

Kaplan Meier Estimation of Survivor Function – <u>fifth and sixth death</u>

				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 x (1-(1/15)) =0.747**.

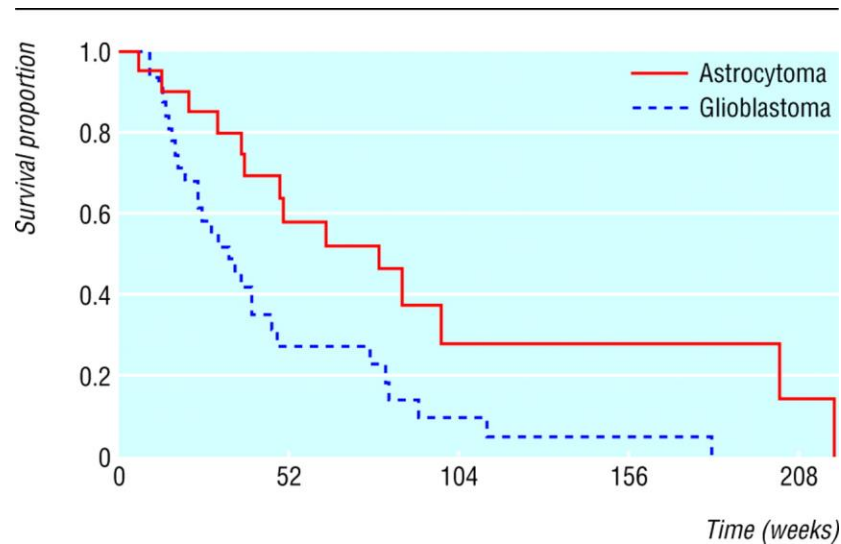| Weeks in follow-up (t)	N at risk at time t	N of deaths at time t	Prob. of death at time t	Prob. of no death at time t	Prob. of surviving up to and including time t
30	17	1	0.059	0.941	0.80
31	16	0	0	1	0.80 x 1 = 0.80
37	15	1	1/15=0.067	0.933	0.80 x 0.933 = 0.747

***

KM plot for 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

***

Survival by cancer type

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

Bland JM, Altman DG. The logrank test. BMJ. 2004 May 1;328(7447):1073. doi: 10.1136/bmj.328.7447.1073. PMID: 15117797; PMCID: PMC403858

Comparing two samples

- Could compare median survival time, or probability of surviving up to any particular time.

- Better to use a test which compares survivors over whole follow-up period.

- **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.

- **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 it has a Chi^2 distribution with 1 degree of freedom.

Log rank test to compare 2 groups

Astro	Death (=1)	Glio	Death (=1)
6	1	10	1
13	1	10	1
21	1	12	1
30	1	13	1
31	0	14	1
37	1	15	1
38	1	16	1
47	0	17	1
49	1	18	1
50	1	20	1
63	1	24	1
79	1	24	1
80	0	25	1
82	0	28	1
82	0	30	1
86	1	33	1
98	1	34	0
149	0	35	1
202	1	37	1
219	1	40	1
=14 deaths		40	1
		40	0
		46	1
		48	1
		70	0
		76	1
		81	1
		82	1
		91	1
		112	1
		181	1
		=28 deaths	

20/51   31/51

Week	Overall Observed Deaths	Expected Deaths – Astro	Expected Deaths – Glio	Observed Remainder – Astro	Observed Remainder – Glio
6	1/51	0.392157	0.607843	19	31
10	2/50	0.76	1.24	19	29
12					
13					
14	(19/50)*2		(31/50)*2		
15			$x^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$		
...					
Total (Expected)		Sum	Sum		

The Log rank test compares observed number of events, say O_i for cancer type i, to the expected number by calculating the test statistic

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$$

This value is compared to a χ^2 distribution with $(g-1)$ degrees of freedom, where g is the number of groups.
In this manner, a P-value may be computed to calculate the statistical significance of the differences between the complete survival curves.
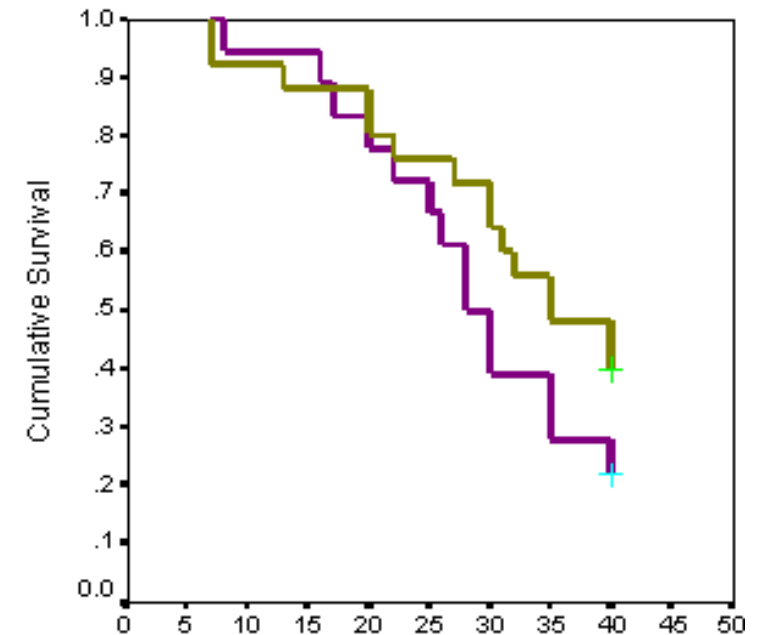
***

Log Rank test

- Unlikely to detect a difference between Groups if survivors cross over during follow-up.

- Assumes non-informative censoring.

- Can be extended to compare more than 2 groups.

- **But**

  - Only provides a p-Value, not an estimate of size of difference between groups or a confidence interval.

***

Hazard Function

- Conceptually, hazard is opposite to survival. It is the event rate or death rate (in case of analysis of impact of diseases on humans). It can also be termed as *failure rate*.
  - hazard is **the probability that an individual who is under observation** at a **time t has an event at that time**.
- A hazard function is represented as h(t).
- Mathematically:

$$h_T(t) = \lim \frac{P(t < T < t + \delta \mid T < t)}{\delta}$$

$$\delta \to 0$$

Hazard Function

- Hazard at time t is potential per unit time for the event to occur given that the subject has survived till time t.

- Basically, it is the **rate** of event at time t.

- It is clear from the expression and definition that *hazard is a rate* rather than being a probability.

- Hazard and survivor functions are just two ways of specifying a survival distribution.
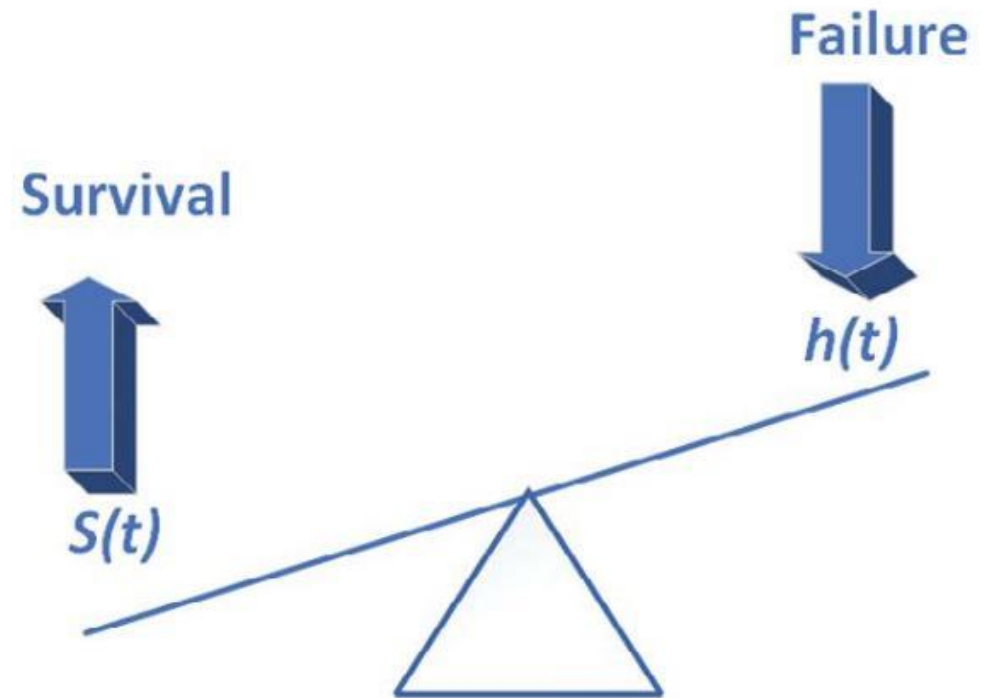
Relationship between survival and hazard

Analysis of relationship between survival and hazard functions

So,

- Survivor function:

  S (t) = P (T > t) = 1 − P (T ≤ t)

- P(T <= t) is also known as cumulative risk function and denoted by F(t).

  F(t) = 1− S(t)

- From theory, we know that probability density function (denoted by f (t)) is the derivative of the distribution function.

$$f(t) = \frac{d}{dt} F(t)$$

$$= \frac{d}{dt} \left[1 - S(t) \right]$$

$$= -\frac{d}{dt} S(t)$$

Analysis of relationships between survival and hazard functions

- Density function is the negative of the derivative of survivor.

- From the previous equation, by the theorem of conditional probability and omitting suffix T, we get the expression of hazard function as:

$$h(t) = \frac{f(t)}{S(t)}$$

- Basically, the hazard at time t is the probability that an event occurs in the very short neighborhood of time t divided by the probability that the subject is alive at time t.

- Hazard being a ratio of two probabilities, it is a rate rather than a probability itself.
    - Hence, we have the **hazard ratio**

Analysis of relationships between survival and hazard functions

- Unfortunately, unlike $S(t)$ there is no simple way to estimate $h(t)$. Instead, a quantity called the *cumulative hazard H(t)* is commonly used.

- Cumulative hazard function is the total hazard from zero to time t.

$$H(t) = \int_0^t h(u)\,du$$

- The formal relationship between cumulative hazard and survival is given by:

$$S(t) = e^{-H(t)}$$

- Hence survival and hazard are both functions of time t.

- It seems like survival probability of a subject is dependent only on time - univariate model [**for now**]

Cox proportional hazards (Cox-PH) model

- Proposed by Cox (1972) - with no assumption on h_0(t).

- It is the most widely used estimator.

- It is easy to implement, takes covariates into account and provides interpretable results.

- It is a semi-parametric method that aims at modeling the hazard function.
  - it makes a ***parametric assumption concerning the effect of the predictors*** on the hazard function, but makes ***no assumption regarding the nature of the hazard function h(t) itself.***

- We saw:

$$h_T(t) = \lim \frac{P(t < T < t + \delta \mid T < t)}{\delta}$$

$$\delta \rightarrow 0$$

Cox Proportional Hazards

- Cox-PH models the hazard function as follows:

Baseline hazard

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n)$$

Hazard ratio

$$h(t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)$$

- The model consists of 2 parts:
  - **the baseline hazard:** it describes how the risk evolves through time.
  - **the hazard ratio:** it models the effect of the explanatory covariates on the risk.

Cox Proportional Hazards

- With this semi-parametric function, the model relies on a strong proportionality assumption:
  - For example, if a patient has a risk of relapse at an initial observation that is twice as low as another patient, then for all subsequent time observations, the risk of relapse remains twice as low.

Baseline hazard

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n)$$

Hazard ratio

Application of Cox-PH model

The outputs of the model are highly interpretable.

At the **instance level**, the model provides for each observation:

- **A risk score**: the higher the risk, it is more likely that a patient will die.
- **A survivor** function: it enables analysts to assess the probability of surviving at least to a point t over time.
- **A hazard function:** it serves the same purpose.

The outputs of the model are highly interpretable.

At the **global level**, the model can be explained through coefficients:

- a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated

- for a positive coefficient, the higher it is, the stronger the impact on the relapse.

V1
V2
V3
...

−0.2   −0.1   0   0.1

Coefficient

Baseline hazard

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n)$$

Hazard ratio
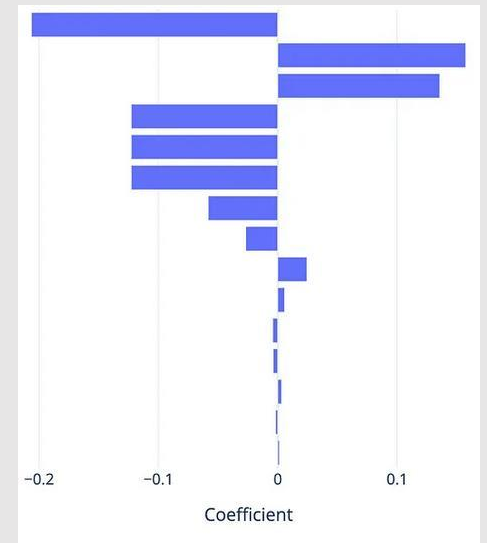
Example of time-to-event data

- Let's consider the [Prostate cancer (TCGA, PanCancer Atlas)](#) - TCGA PRAD

- 494 samples of men with Prostate Adenocarcinoma
  - Gene Expression Data
  - Mutation Data
  - Protein Expression data

Survival Analysis of TCGA PRAD

The androgen receptor (AR) is a member of hormonal transcription factors.

The expression of AR protein and its activation by male hormone androgen are fundamental to prostate development during pubertal and malignant transformation during later ages.

19 genes downregulated by Androgen Receptor in PCa

BCHE, CDK8, **CTBP1**, ACKR3, DDC, DPH, N1, **HES6**, **MMP16**, MYC, PEG3, **PIK3R3**, **PRKD1**, **SCNN1A**, SDC4, SERPINI1, SLC29A1, ST7, TULP4

Progression-free survival analysis of TCGA PRAD

- Note: Disease-free survival (DFS) is defined as **the length of time after primary treatment for a cancer ends that the patient survives without any signs or symptoms of that cancer**

- At 50.73 months:
  - 87.87% event-free survival in un-altered group of patients
  - 70.86% event-free survival in altered group of patients

Logrank Test P-Value: 0.0267

Disease Free
- ■ Altered group
- ■ Unaltered group
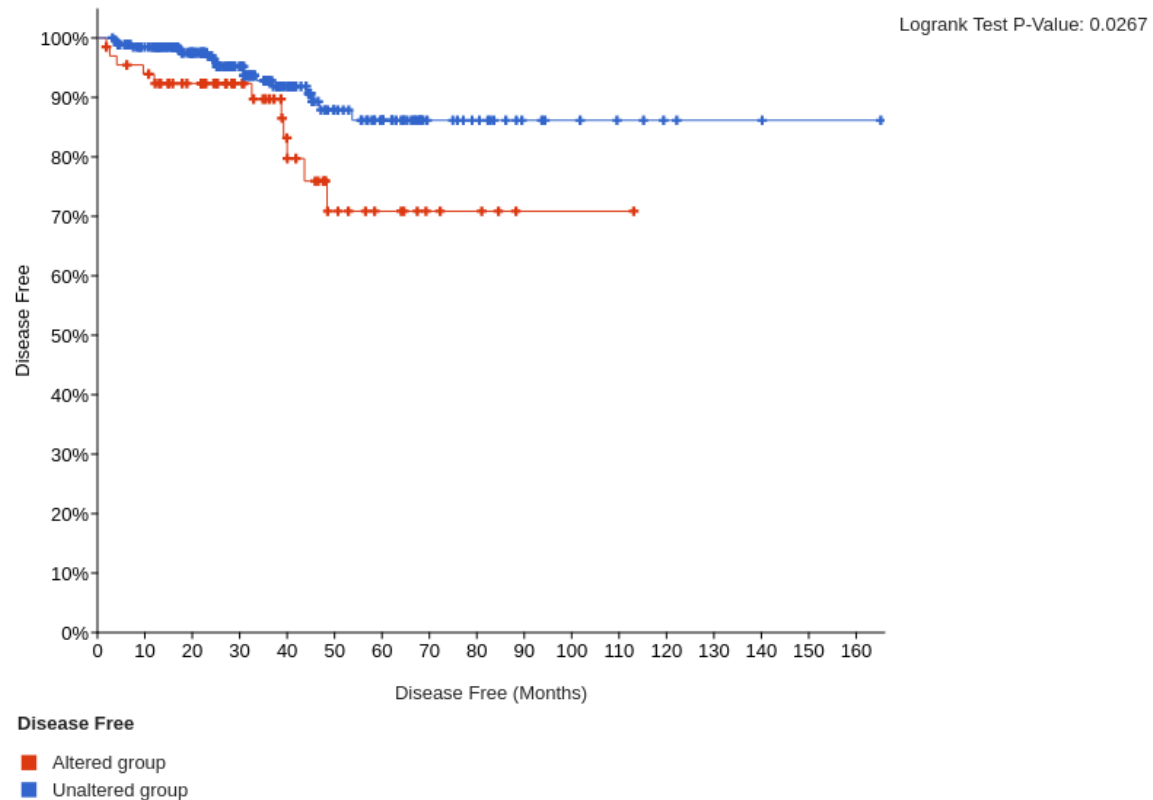
Analysis done on the CBioPortal site: www.cbioportal.org

How Is Machine Learning Leveraged for Survival Analysis?

- **Cox-PH model**

- **Random Survival Forest model**
  - Just as standard Random Forests, Random Survival Forest, in a training, consists of a number of survival trees on various sub-samples (often drawn with replacement) of the dataset and using averaging for more accurate prediction and limited over-fitting.
  - The main difference lies in the metric used to assess the quality of a split: **log-rank** which is typically used when comparing survival curves among two or more groups.

How Is Machine Learning Leveraged for Survival Analysis?

- **Gradient Boosting Survival model**
  - Gradient Boosting when applied to survival analysis is also very similar.
    - it consists of combining in an additive manner, the predictions of multiple base learners to obtain a powerful overall model.
    - The base learners, also called weak learners, are often very simple models.
    - It differs from Random Forests as the survival trees are not trained independently but sequentially in a greedy stagewise fashion.

How Is Machine Learning Leveraged for Survival Analysis?

- **Survival Support Vector Machine model**
  - Survival Support Vector Machine (SVM) can also be extended to survival analysis.
    - It is also a very versatile model as it can account for complex, non-linear relationships between features and survival via the so-called kernel trick.
  - However, its predictions cannot be easily related to the standard quantities of survival analysis, that is, the survivor and the cumulative hazard function.

Comparison of different survival models

- **Concordance Index**
  - The most frequently used evaluation metric is the concordance index, also referred to as the c-index.
  - It measures the capability of the model to provide a reliable ranking of survival times based on individual risk scores.
  - It is computed as the proportion of concordant pairs in a dataset.
- More specifically, let's consider two observations *(i,j)*:
  - First, to be comparable, the observation with lower time needs to have experienced the event.
  - Second, if comparable, it is concordant if the risk estimated by the survival model is higher for individuals with shorter survival times.
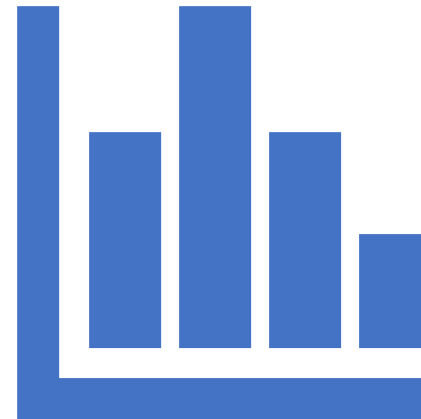
Comparison of survival models

- **Cumulative/dynamic AUC**
  - ROC curve is extended to censored survival times.
  - The idea is to consider several points in time.
  - At each point, we consider separately:
    - the cumulative cases: all individuals who have experienced the event before or at time t.
    - the dynamic controls: individuals who will experience an event after time t.
  - We can then evaluate the model on its ability to distinguish subjects who will experience an event over time (sensitivity) from those who will not (specificity).

Finally, some key requirements for the analysis of survival data

- **Uninformative censoring**
  - Censored individuals, because of loss to follow-up at a given point in time, should be as likely to have a subsequent event as those individuals who remain in the study

- **Length of follow-up**
  - Time to event studies must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests.

- **Completeness of follow-up**
  - Each patient who does not have an event can be included in a survival analysis for the period up to the time at which they are censored, but completeness of follow-up is still important.
  - In general, disparities in follow-up caused by differential drop-out between arms of a trial or different subgroups in a cohort study need to be investigated.

- **Cohort effect on survival**
  - In survival analysis, there is an assumption of homogeneity of treatment and other factors during the follow-up period.
  - But, in a long-term observational study of patients of cancer, the case mix may change over the period of recruitment, or there may be an innovation in ancillary treatment.

- **Between-centre differences**
  - In a multicentre study, it is important that there is a consistency between the study methods in each centre.

Survival Analysis in action

- In Python we can use 2 libraries:
  - scikit-survival
  - Lifelines

- Example Survival Analysis using Python
  - **Haberman Dataset**: The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.
  - Link to data: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set/download?datasetVersionNumber=1
  - Link to Google Collab notebook: https://drive.google.com/file/d/1lOEdnc9t2H02Z7omqZjuhkpjmXb9Dgd_/view?usp=sharing

Resources used to develop the slides

The slides have been created with materials from web resources and:

1. Book - Survival Analysis with Python

2. CBioPortal Resource for data and survival analysis of TCGA Prostate Cancer

3. Kaggle for datasets for the demo of survival analysis

4. Bland JM, Altman DG. The logrank test. BMJ. 2004 May 1;328(7447):1073. doi: 10.1136/bmj.328.7447.1073. PMID: 15117797; PMCID: PMC403858 for the sample data to derive the survivor and hazard function

5. A few slides have been taken from a presentation on Survival analysis by Anne Segonds-Pichon of the Baraham Institute (***)

Interesting articles on survival analysis

- Clark TG, Bradburn MJ, Love SB, Altman DG. **Survival analysis part I**: basic concepts and first analyses. Br J Cancer. 2003 Jul 21;89(2):232-8. doi: 10.1038/sj.bjc.6601118. PMID: 12865907; PMCID: PMC2394262.

- In J, Lee DK. **Survival analysis: part II** - applied clinical data analysis. Korean J Anesthesiol. 2019 Oct;72(5):441-457. doi: 10.4097/kja.19183. Epub 2019 May 17. Erratum in: Korean J Anesthesiol. 2023 Feb;76(1):84-85. PMID: 31096731; PMCID: PMC6781220.

- Bradburn, M., Clark, T., Love, S. *et al.* **Survival Analysis Part III**: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer* **89**, 605–611 (2003). https://doi.org/10.1038/sj.bjc.6601120

- Clark, T., Bradburn, M., Love, S. *et al.* **Survival Analysis Part IV**: Further concepts and methods in survival analysis. *Br J Cancer* **89**, 781–786 (2003). https://doi.org/10.1038/sj.bjc.6601117

- Rai, S., Mishra, P. & Ghoshal, U.C. **Survival analysis: A primer for the clinician scientists**. *Indian J Gastroenterol* **40**, 541–549 (2021). https://doi.org/10.1007/s12664-021-01232-1

Acknowledgements

Thank You very much, for your attention!!!