What is Kaplan Meier curve?

* Graph plot of the survival probability against time-to-events, showing the probability of a subject will survive up to time
* Non-parametric estimate which does not make assumption about the distribution of data
* Estimate from data that could be censored, do not experience the events

Censored data (the data is incomplete without end time for events)

* Not experienced the event before the end of Kaplan Meier analysis
* Follow up is missing or withdrawn from the analysis

If do not include the censored data, the sample size of the study may become small

For Kaplan Meier curve

In specific question:

What is the difference in survival of **Malignant neoplasm of breast (disorder)** treated with **Paclitaxel 100 MG Injection** compared to **100 ML Epirubicin Hydrochloride 2 MG/ML Injection**?

Here the event in interest is the death of the patient, we assume the death is caused by the cancer.

1. Get the time range for the Kaplan Meier analysis. (Here let start time be the earliest time for a patient got treatment and end time as the time the data extracted)
2. Clinical outcome (alive or death) is used to get status, if death then status as 1 otherwise status as 0
3. Get start time the patient got treatment here is the earliest time the patient got specified medication(data must have start time)
4. Get end time from patient table (DEATHDATE), if alive then the end time is the end time of the Kaplan Meier analysis,
5. Then compute duration and get the data with columns: duration, status, other group variables to do plot.

|  |  |  |
| --- | --- | --- |
| Duration/Time | Status/events | group |
|  |  |  |

1. And then for each time point for a specific group, we count the number of events censored whose status value is 0; and the number of events occurred which are events with status 1(patient died); The number of events at risk is equal to the number of events at risk at last time point minus the number of events censored minus and the number of events occurred

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time(t) | Number of event censored() | Number of event occurred() | Number of event at risk() | Survival probability( ) |
|  |  |  |  |  |

According to the formula to calculate the survival probability

= where, = - -

|  |  |  |
| --- | --- | --- |
| Time(t) | Survival probability() | group |
|  |  |  |

Reference:

<https://yuzar-blog.netlify.app/posts/2021-01-03-survival-analysis-1-a-gentle-introduction-into-kaplan-meier-curves/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/>

<https://towardsdatascience.com/kaplan-meier-curve-explained-9c78a681faca>

What is log-rank Test?

* Hypothesis test to test whether there is a statistical difference between groups in the survival probability of events at any time. The null hypothesis is that there is no statistical difference between groups hence if the pvalue is smaller than the significance level(normally 0.05), we reject the null hypothesis which means there is statistical difference between groups.

|  |  |  |
| --- | --- | --- |
| Time(t) | Status/events | group |
|  |  |  |

To manually calculate the pvalue with two groups, we firstly get a table, where for each time to event, where m is the number of subject that experienced events at that time; q is the number of censored subjects at that time; n is the number of subjects at risk at that time(= the number of subjects at risk last time minus the number of subject event happened and censored subject); e is the number of expected observed number of event happened

= =

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time(t) | m1 | q1 | n1 | m2 | q2 | n2 | e1 | e2 |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

The Log-statistics = where ,is the sum of the across time which represent the difference between the number of observed events and the number of expected events in the second group.

Then the pvalue extracted from the log statistic to compare with a chi square distribution with degrees of freedom which is the number of groups minus 1.

There are package in R and python can calculate the pvalue with data provided with time, status, group.

Cox proportional hazard model

Limitation of logrank test:

* only analyze the effect of one variable
* does not allow us to investigate the simultaneous impact of multiple categorical variables or continuous variables
* log-rank test can only tell us if there is a statistically significant difference between groups. It cannot provide a hazard rate or hazard ratio. Hence, it cannot quantify this difference

Assumption for Cox

* The proportional hazards assumption: In CPH, the hazard ratio is assumed to remain constant throughout the follow-up
* Independence of survival times between distinct individuals in the study population
* censoring is uninformative about the outcome of interest, censored have the same risk of suffering the study end-point as those who continue to be followed
* **It is a semi-parametric model** because there are no assumptions about the shape of the baseline hazard function

Formula for the model:

Logo, company name

Description automatically generated

hazard ratio (HR)

* It is the ratio of the total number of observed to expected events in two independent comparison groups

e.g. Diagram

Description automatically generated

* The estimated coefficients in the Cox proportional hazards regression model represent the change in the expected log of the hazard ratio relative to a one unit change in X1, holding all other predictors constant.
* As the time component is canceled in the ratio, HR does not depend on time t, indicating a proportional hazard over time.
* The associations are quantified by the regression coefficients (b1, b2, ..., bp)

Reference:

<https://www.nature.com/articles/6601118>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7876211/>

<https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival6.html>

<https://www.r-bloggers.com/2019/09/survival-analysis-with-strata-clusters-frailties-and-competing-risks-in-in-finalfit/>

<https://journals.lww.com/anesthesia-analgesia/Fulltext/2021/04000/Kaplan_Meier_Curves,_Log_Rank_Tests,_and_Cox.7.aspx>