Iciar Luna Vázquez

SURVIVAL ANALYSIS

In this blog entry survival statistical analysis will be discussed, and the main examples used for the explanation of this type of analysis will mainly revolve around oncological studies.

Introduction to survival analyses

The survival analysis can be defined as the collection of procedures for data analysis where the outcome variable is the time until an event of interest occurs. This event, in most cases is death and is referred to as **survival time.** This, however, can also be applied to other events such as time until first metastasis, or time to relapse from complete remission, among others.

❖ Censoring data

The main difficulty encountered when performing survival analysis appear from the fact that some patients haven't experienced the event of interest, therefore, their survival time is unknown, this is known as **censoring.** We can encounter this phenomenon under different situations:

1. The patient hasn't experienced the outcome we are studying, for instance death, or first metastasis appearance.
2. The patient is lost during the study period hence, follow-up cannot be done.
3. The patient experiences a different event, for instance death from a different reason from that of our experiment, making the follow-up process impossible.

By censoring data, we are underestimating the true, but unknown, time to event. There are different types of censoring, but the most common ones are:

- **right censoring**, which allows us to visualise the survival process of an individual despite it being "incomplete", in other words, the event hasn't yet occurred, and it is only known that it may occur after the end of the study.
- **left censoring**, which is the opposite situation, which indicates subject has been lost or dropped out of the study, before the endo of the study.

To explain this, we will take the following example: where the time to breast cancer recurrence is studied in a series of patients who have undergone surgical removal of the tumour.

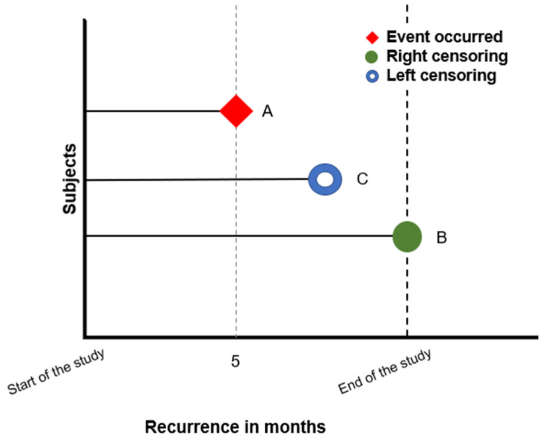


Figure 1. Illustrates the recurrence of breast cancer (event) in patients who have undergone surgical removal of the tumour, (start of study).

Subject A (event occurred) represents a patient that was examined 5 months after the start of the study and the cancer had return. On the contrary, subject B (right censoring), represents a patient that reached the end of the study without disease recurrence, this means that the actual time of the event is only known to be at some point after the end of the study. Finally, subject C (left censoring), represents a patient that has been lost, or dropped out of the study.

Survival function

The survival function (S(t)), is defined as the probability that a patient (in the case of cancer studies), device, etc. will survive past a certain period of time (t), in other words, that our event of interest hasn't occurred at time t. It is computed as

$$\mathbf{S(t) = Pr\ (T>t)}$$

It is very important to understand, that at time zero (t=0), all patients who entered the study are alive, hence, survival is of 100%, every time a patient dies, the survival probability decreases, hence why if the study were to be prolonged to infinity (t= ∞) the survival probability would drop to 0.

The **Kaplan-Meier estimator** is the most common method used to estimate survival function, and it is frequently used in clinical research to estimate the proportion of subjects alive at specified time points after treatment. That is, with this method, we are estimating the survival probability each time an event occurs, and it also allows us to compare two or more groups of subjects, once the survival curves have been generated.

$$\hat{\boldsymbol{S}}(\boldsymbol{t}) = \prod_{\{i:t_i \le t\}} \left[\mathbf{1} - \frac{\boldsymbol{d_i}}{\boldsymbol{n_i}}\right]$$

$$where\ d_i\ is\ the\ number\ of\ events, n_i\ the\ total\ individuals\ at\ risk$$

To understand how to estimate the survival probability using the Kaplan-Meier approach, also known as the product limit estimate, we will use the following the first example.

i.e. 100 women with breast cancer, underwent a surgical procedure to remove the tumour, they were then checked every 2 months for 6 months for possible recurrence. During the first follow-up (month 2) 3 women left the study, and 5 had recurrence; during the second follow-up (month 4) 4 women died form external causes, and 1 woman had relapse; during the last follow-up (month 6) 2 women disappeared form the study, and 5 women had relapse.

Time period	At risk	Censored	Recurrence	Survived	Survival probability
Start study	100			100	1
2nd month	100	3	5	95	1*(95/100) = **0.95**
4th month	95-3 = **92**	4	1	92-1= **91**	0.95*(91/92) = **0.93**
6th month	91-4 = **87**	2	5	87-5 = **82**	0.93*(82/87) = **0.88**

Table 1. Imaginary data for Kaplan-Meier analysis.

If we were to represent these survival probability estimates, we would obtain a survival curve were the symbols would represent each event time, whether there is a relapse or censoring. With the survival curve we could estimate the probability a participant has of surviving past a specific time point, known as the **X-time survival**; as well as the **median survival time** which indicates the time at which half the participants have experienced the events.

- **Standard error**

The standard error (SE) for the Kaplan Meier estimate can be calculated as follows:

$$SE(S_t) = S_t \cdot \sqrt{\sum \frac{D_t}{N_t \cdot (N_t - D_t)}}$$

$$\textit{where } D_t \textit{ is the number of events}, N_t \textit{ the total individuals at risk}$$

an example can be seen in Table 2.

- **Confidence intervals (CI)**

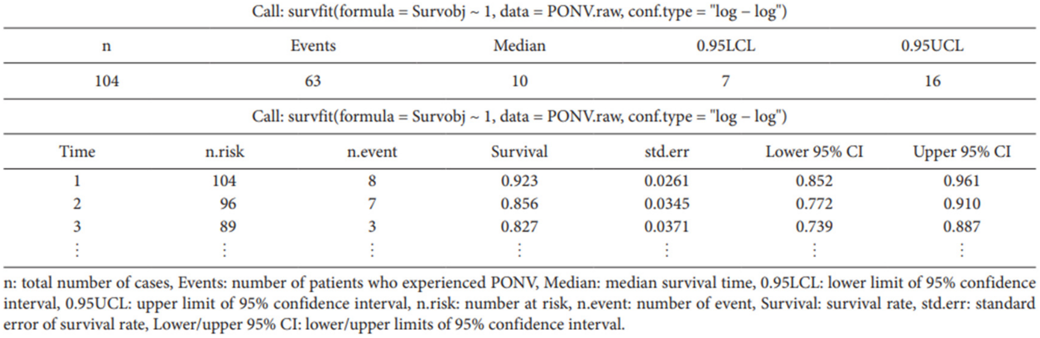
It is important that we understand when calculating CI at any particular time, that they are asymmetrical as the survival percentage cannot possibly go below 0.0% nor above 100%, the reason for this is that survival probability cannot go above a 100 nor below 0.

These lower and upper limits are calculated as:

$$S(t) \in \left(\left[\hat{S}(t) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}(t)}{n}} \right], \left[\hat{S}(t) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}(t)}{n}} \right] \right)$$

$$\textit{where } Z_{1-\frac{\alpha}{2}} \textit{ is the } 1 - \frac{\alpha}{2} \textit{th quantile of standard normal distribution}$$

We can see some examples for this calculation in Table 2, taken from (In and Lee 2019).



Call: survfit(formula = Survobj ~ 1, data = PONV.raw, conf.type = "log – log")

n	Events	Median	0.95LCL	0.95UCL
104	63	10	7	16

Call: survfit(formula = Survobj ~ 1, data = PONV.raw, conf.type = "log – log")

Time	n.risk	n.event	Survival	std.err	Lower 95% CI	Upper 95% CI
1	104	8	0.923	0.0261	0.852	0.961
2	96	7	0.856	0.0345	0.772	0.910
3	89	3	0.827	0.0371	0.739	0.887
⋮	⋮	⋮	⋮	⋮	⋮	⋮

n: total number of cases, Events: number of patients who experienced PONV, Median: median survival time, 0.95LCL: lower limit of 95% confidence interval, 0.95UCL: upper limit of 95% confidence interval, n.risk: number at risk, n.event: number of event, Survival: survival rate, std.err: standard error of survival rate, Lower/upper 95% CI: lower/upper limits of 95% confidence interval.

Table 2. Kaplan-Meier estimations and survival table. Where the first onset time of post-operative nausea and vomiting was analysed.

<u>Hazard ratio</u>

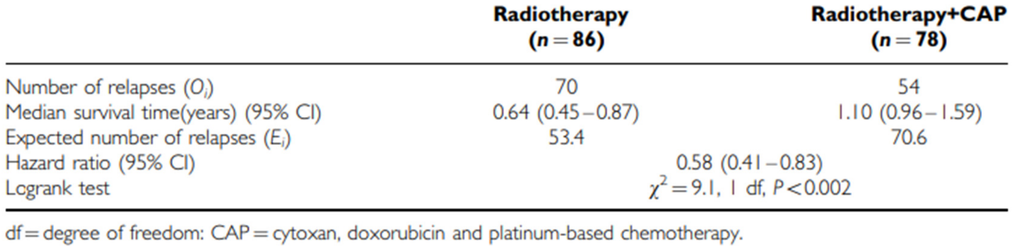
The hazard ratio is defined as the relative risk of an event (e.g., disease progression) occurring in one trial group compared to the other, over the entire duration of the study. This is commonly used for the comparison of survival rate for a new drug compared to the standard of care. Let's say, for instance, the administration of a new oncological drug targeting specific oncogenes for breast cancer, compared to the standard chemotherapy.

$$\boldsymbol{HR = \frac{O_1/E_1}{O_2/E_2}}$$

$where\ O_1\ is\ the\ observed\ event\ rate\ and\ E_1\ the\ expected\ event\ rate\ for\ the\ new\ drug;$

$O_2\ is\ the\ observed\ event\ rate\ and\ E_2\ the\ expected\ event\ rate\ for\ standard\ chemotherapy$

An example for HR calculation, taken from (Clark et al. 2003), is shown in Table 3, that shows the differences in (relapse-free) survival in the lung cancer trial.



	Radiotherapy ($n = 86$)	**Radiotherapy+CAP ($n = 78$)**
Number of relapses (O_i)	70	54
Median survival time(years) (95% CI)	0.64 (0.45–0.87)	1.10 (0.96–1.59)
Expected number of relapses (E_i)	53.4	70.6
Hazard ratio (95% CI)	0.58 (0.41–0.83)	
Lcgrank test	$\chi^2 = 9.1$, 1 df, $P < 0.002$	

df = degree of freedom: CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

Table 3. Differences in (relapse-free) survival in the lung cancer trial, where radiotherapy with and without adjuvant chemotherapy (CAP) was compared.

The HR allows us to calculate the percentage of progression or death risk reduction. This percentage can be calculated as 1-HR. Therefore, lower the HR, the higher the percentage of risk reduction. We can interpret the results as follows:

- HR=2, there is twofold risk of an event occurring.
- HR=1, there is no differences in survival between both groups, hence the new oncological drug is no different from the chemotherapy.
- HR = 0.5 indicates there is half the risk of an event occurring in one group relative to the other, thus there is a 50% risk reduction.

For instance, in the treatment of Breast cancer with a new specific target drug compared to standard chemotherapy, a HR of 0.73 for the overall survival means that there is a reduction of risk of progression/death of 27%. Thus, the new drug has a 27% higher chance of reducing risk of relapse or death compared to standard chemotherapy.

- ❖ **Comparing survival**

When the HR is considered constant, we can use the **log-rank test** to calculate the expected number of events in each group allowing us to determine whether the difference between the survival time from each group are statistically significant or not.

$$\boldsymbol{Log - rank\ test = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}}$$

We can calculate the number of expected events as a $Ei = \frac{Ni}{Oi/Nt}$

The test statistic and the significance of this test can be drawn by comparing the calculated value with the critical value, using the Chi-square table. When the test statistic value is less than the critical value for degree of freedom equal to one, we can say there is no difference between the two groups regarding the survival.

i.e. We want to compare the combination of treatments in patients with lung cancer, where group 1 will receive chemotherapy before surgery and group 2 will receive chemotherapy after surgery, and we will measure for death, were participants will be measured for 2 years.

Table 4. Represents imaginary data generated for the 2 groups. Where: N_t, is the total number at risk, N_b is the number at risk before surgery and N_a after surgery, D_a and D_b is the number of deaths in each group and C_a and C_b is the censored number in each group. St represents the survival probability.

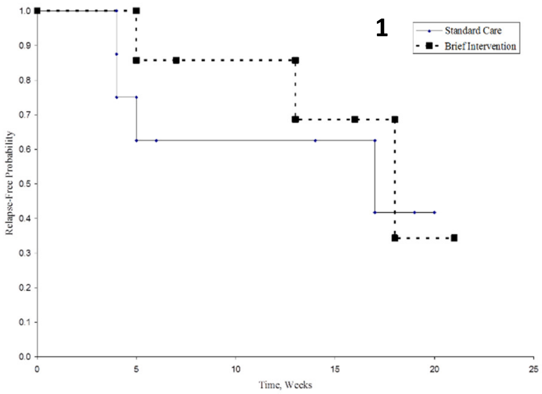
Chemotherapy before surgery					Chemotherapy after surgery					TOTAL	
Time	N_b	C_b	D_b	St_b	Time	N_a	C_a	D_a	St_a	N_t	O_t
0	10			1.00	0	10		0	1.00		
8	10	1	1	0.90	8	10		0	1.00	20	1
12	8		1	0.79	12	10		1	0.90	18	2
14	7	1	0	0.79	14	9	1	1	0.80	16	1
20	6	1	1	0.66	20	7	1	1	0.69	13	2
24	4		1	0.49	24	5		1	0.55	9	2

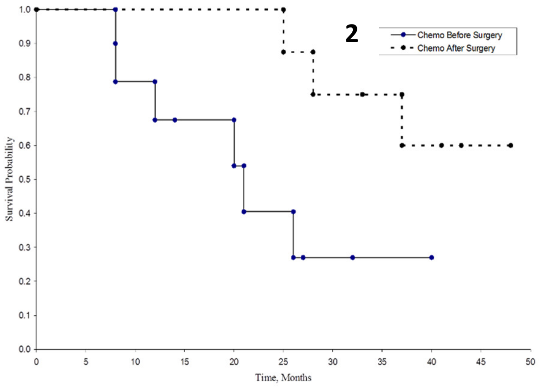
Chemotherapy before surgery				Chemotherapy after surgery			
Time	N_b	O_b	E_b	Time	N_a	O_a	E_a
8	10	1	0.50	8	10	0	0.5
12	7	1	0.89	12	10	1	1.11
14	5	0	0.44	14	8	1	0.56
20	4	1	0.92	20	5	1	1.08
24	2	1	0.89	24	3	1	1.11
Total		4	3.64			4	4.36

$$Log\ rank\ test = \frac{(4-3.64)^2}{3.64} + \frac{(4-4.36)^2}{4.36} = 0.036 + 0.026 = 0.067$$

If we check the Chi-square table with 1 degree of freedom and compare our calculated value (0.067), with the critical value, we will be able to determine if we can or not reject the H_0. The critical value for the rejection of the H_0 is $X^2 > 3.84$. In our case, we are not able to reject the H_0 as X^2=0.067.

The figure below represents the survival curve comparison between two groups that have very little differences between one another (1), a Log-rank test value below 3.84 and the comparison between two groups that have significant differences (2) between one another Log-rank test value above 3.84. (LaMorte 2016)





Conclusion

Survival analysis allows us to estimate the time to event probability, it is necessary to censor data when performing these analysis as if not we could be altering the results of our study. To properly estimate this survival probability, we can apply/use the Kaplan-Meier approach. With this approach we can also evaluate the HR which represents the percentage risk of event occurring reduction. We could use this statistical parameter to compare two groups, whenever it is considered to be constant, using the Log-rank test whenever.

USEFUL INFORMATION

Additional information and webpages to understand the Survival Analysis

- Survival Analysis, Boston University School of Public Health (LaMorte 2016)
- Survival Analysis Part I: Basic concepts and first analyses. Tutorial Paper. (Clark et al. 2003)
- Survival analysis: part II – applied clinical data analysis. (In and Lee 2019)
- Hazard Ratio in Clinical Trials (Spruance et al. 2004)
- Censored Data Analysis exercises
- The Basics of Survival Analysis (Chun Chan 2016)
- An Introduction to Survival Analysis for Clinical Trials (Toupin 2017)
- Confidence Interval for the Survival Function
- Hazard Ratio calculation
- H. Motulsky. Intuitive Biostatistics: A Nonmathematical Guide to StatisticalThinking, 3rd edition. Oxford University Press (2013)
- DATAtab, Survival Analysis [Simply Explained], YouTube video. https://youtu.be/Wo9RNcHM_bs