University of Waterloo

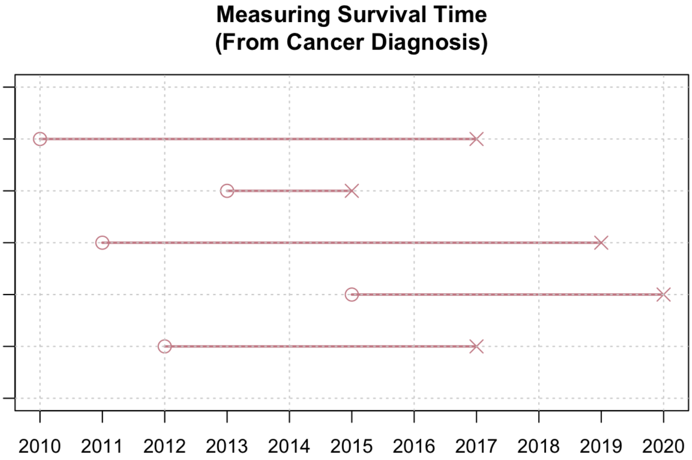
An Introduction to Survival Analysis

Alina Hu, Calista Kurniawan, and Maya Le
Mentor: Xianwei Li

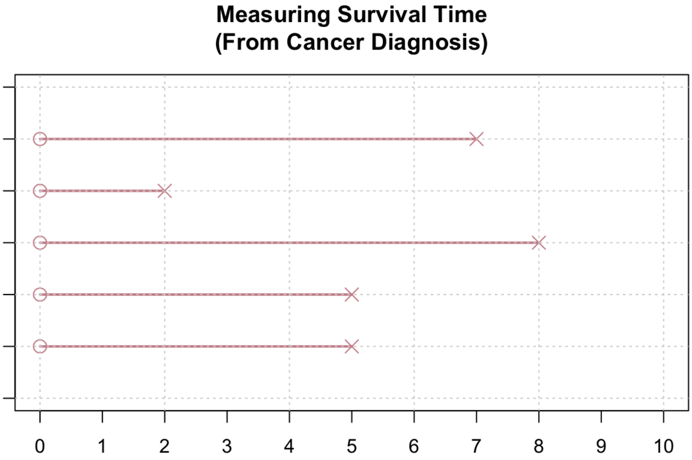
WiM Direct Reading Program, Fall 2024

Outline

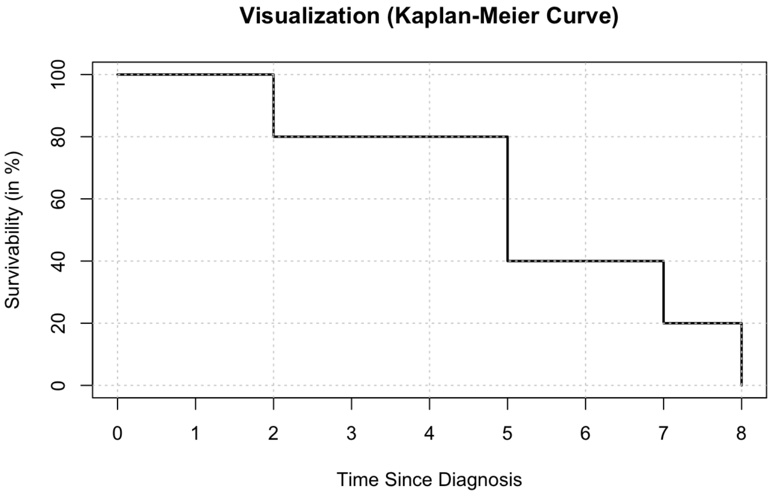
Intuition


Measuring Survival Time
(From Cancer Diagnosis)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

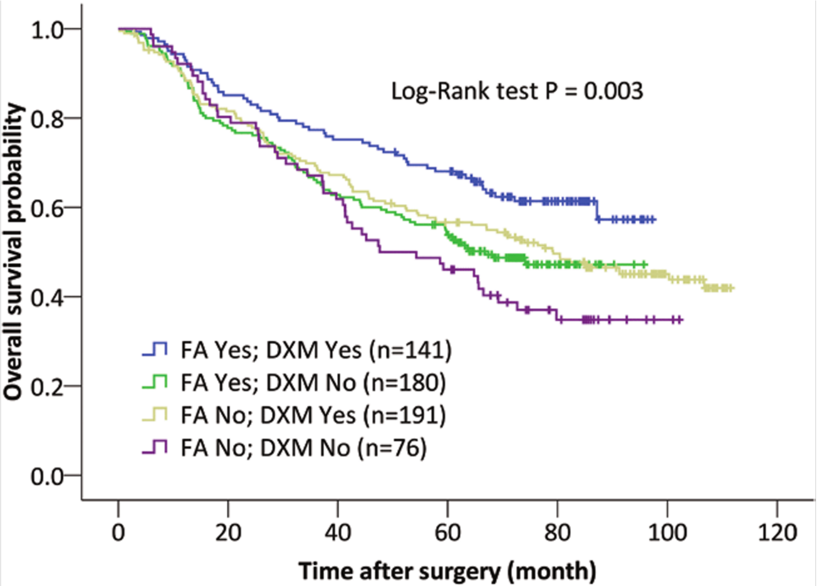
Intuition


Measuring Survival Time
(From Cancer Diagnosis)
0
1
2
3
4
5
6
7
8
9
10

Intuition


Visualization (Kaplan-Meier Curve)
Survivability (in %)
0
20
40
60
80
100
0
1
2
3
4
5
6
7
8
Time Since Diagnosis

Intuition


Log-Rank test P = 0.003
FA Yes; DXM Yes (n=141)
FA Yes; DXM No (n=180)
FA No; DXM Yes (n=191)
FA No; DXM No (n=76)
Overall survival probability
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Time after surgery (month)

Introduction to Survival Analysis

Definition

- **Survival analysis** is a branch of statistics that measures the time until an event occurs.
- **Survival time** is the particular variable of interest.
  - Exposure $\rightarrow$ Event
  - Ex. Time of cancer diagnosis to death
- Survival Analysis doesn't have to just be involved with death, but in the same lens of cancer, it could be the time of complete remission to relapse

Censoring

Definition

- Censoring occurs when we don't know the exact time to event.
- We don't delete these observations
  - Make a note that the result was censored.
- Different types of censoring
  - Right censoring
  - Left Censoring
  - Interval censoring

Right Censoring

Definition

- Time to the event is GREATER than some value x
  - $t_i > x$
- Study: Estimating survival time after diagnosis of pancreatic cancer (Wahutu, 2016)
  - Consider: Patients still alive at the end of the study; Patients who are lost to follow up

Interval Censoring

Definition

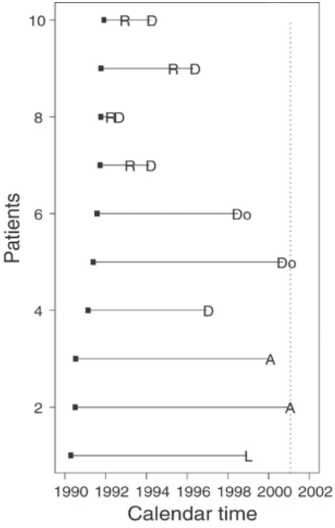
- Time to event is BETWEEN 2 values x_1 and x_2
  - $x_1 < t_i < x_2$
- Study: Oral lesion occurrence in immunosuppressed children (Rodrigues, 2018)
  - Consider: Lesion occurrence is identified by a specialist at regular checkups

Left Censoring

Definition

- Time to the event is LESS than some value x
  - $t_i < x$
- Study: Age at menarche cohort study (Wohlfahrt-Veje, 2016)
  - Consider: Young women enrolled in the study who have already begun menstruating

Intuition


Patients
10
8
6
4
2
R D
RD
R D
Do
Do
D
A
A
L
1990 1992 1994 1996 1998 2000 2002
Calendar time
Patients
10
8
6
4
2
0 2 4 6 8 10 12
Time from diagnosis (years)

Important Functions

Functions of interest

- Survival function: $S(t) = P(T > t)$
- Hazard function: $h(t)$ represents the instantaneous risk of occurrence of the event given the history

Kaplan-Meier Survival Curves

Purpose

A useful non-parametric way to estimate the survival function. We calculating using the formula

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_j}\right)$$

Where d_j is the number of deaths at time t_j and n_j is the number of subjects at risk.

Assumptions

1. Random censoring
2. Non-informative censoring
3. Independence of censoring
4. Survival probabilities do not change over time
5. No competing risks

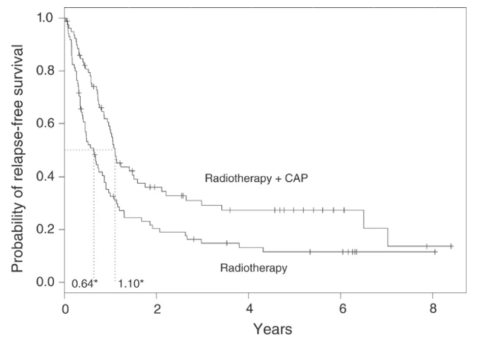
Example: Calculating KM Curves By Hand for Lung Cancer Trial

Figure 1

Radiotherapy (*n* =86)		Radiotherapy+CAP (*n* =78)	
Survival times (days)	**Kaplan–Meier survivor function *S*(*t*)**	**Survival times (days)**	**Kaplan–Meier survivor function *S*(*t*)**
18	1 × (1-1/86)=0.988	9	1 × (1-1/78)=0.987
23[a]	S(18) × (1-0/85)=0.988	22	S(18) × (1-1/77)=0.974
25	S(23) × (1-1/84)=0.977	35	S(22) × (1-1/76)=0.962
27	S(25) × (1-1/83)=0.965	53	S(35) × (1-1/75)=0.949
28	S(27) × (1-1/82)=0.953	76	S(53) × (1-1/74)=0.936
30	S(28) × (1-1/81)=0.941	81	S(76) × (1-1/73)=0.923
36	S(30) × (1-1/80)=0.930	94	S(81) × (1-1/72)=0.910
45	S(36) × (1-1/79)=0.918	97	S(94) × (1-1/71)=0.897
55	S(45) × (1-1/78)=0.906	103	S(97) × (1-1/70)=0.885
56	S(55) × (1-1/77)=0.894	114	S(103) × (1-1/69)=0.872
57	S(56) × (1-3/76)=0.859	115	S(114) × (1-1/68)=0.859
57	S(56) × (1-3/76)=0.859	121[a]	S(115) × (1-0/67)=0.859
57	S(56) × (1-3/76)=0.859	126	S(121) × (1-1/66)=0.846

Example: Comparing KM Curves

Figure 2



Observations

- Overall, Radiotherapy+CAP has a higher survival probability
- The Radiotherapy+CAP group has greater median survival time

Log-Rank Test

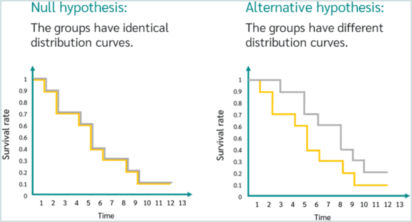
Purpose

A non-parametric test statistic used to compare two survival curves (independent from each other) by calculating

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

where O_i is the observed number of events and E_i is the total expected number of events in each group i.

Hypothesis



Back to Cancer Research Example

Log Rank Test for 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)	
Logrank test	χ^2=9.1, 1 df, P<0.002	

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

Observations

- Log rank test yields a χ^2 value of 9.1 on 1 degree of freedom (P<0.002)
- Hazard Ratio of 0.58 indicates that there is 42% less risk of relapse at any point in time among patients surviving in the combination treatment group compared to those treated with radiotherapy alone
- Indication is present that the combination treatment is more effective than radiotherapy treatment

The Cox Propotional Hazard model

Definition

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

$$\mathbf{X} = (X_1, X_2, \ldots, X_p) \quad \text{explanatory/predictor variables}$$

An expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by **X**.

Product of Two Quantities

$$h_0(t) \times e^{\sum_{i=1}^{p} \beta_i X_i}$$

$h_0(t)$	$e^{\sum_{i=1}^{p} \beta_i X_i}$
Baseline hazard Involves t but not X's	Exponential Involves X's but not t (X's are time-independent)

Semi-Parametric Nature

What does semi-parametric mean?

- Combines parametric and non-parametric components.
- The baseline hazard, $h_0(t)$, is an unspecified function (non-parametric).
- The relationship between the covariates and the hazard rate is expressed parametrically.

Why is this important?

- The Cox PH model is a "robust" model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.
- This property makes the Cox PH Model more flexible than fully parametric models while still allowing meaningful interpretation.

Interpretation

Hazard Ratio (HR)

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})}$$

Measures the relative risk of an event for different covariate levels.

Interval Estimation

- Large sample 95% confidence interval:

$$\exp\left[\hat{\beta}_1 \pm 1.96\sqrt{\mathsf{Var}(\hat{\beta}_1)}\right]$$

where

$$s_{\hat{\beta}_1} = \sqrt{\mathsf{Var}(\hat{\beta}_1)}$$

Cox PH Model Ovarian Dataset Example

Table 1 Hazard ratios from the Cox PH model for the ovarian dataset

From: Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods

	Univariate analysis				Multivariate analysis			
Covariate	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**	**Coefficient (b_i)**	**HR [exp(b_i)]**	**95% CI**	***P*-value**
FIGO stage	0.809	2.24	(2.03–2.48)	<0.001	0.731	2.08	(1.82–2.37)	<0.001
Histology				<0.001				<0.001
Serous	(0.000)	(1.00)			(0.000)	(1.00)		
Mucinous	−0.727	0.48	(0.38–0.61)		−0.422	0.66	(0.50–0.85)	
Endometroid	−1.162	0.31	(0.22–0.45)		0.198	1.22	(0.80–1.85)	
Clear cell	−0.343	0.71	(0.52–0.97)		0.342	1.41	(0.99–2.00)	
Adenocarcinoma	0.119	1.13	(0.74–1.72)		0.501	1.65	(0.91–2.99)	
Undifferentiated	0.390	1.48	(0.81–2.70)		0.746	2.11	(1.03–4.29)	
Mixed mesodermal	0.614	1.85	(1.28–2.66)		0.789	2.20	(1.45–3.35)	
Grade				<0.001				<0.001
1	(0.000)	(1.00)			(0.000)	(1.00)		
2	1.116	3.05	(1.90–4.91)		0.885	2.42	(1.40–4.19)	
3	1.650	5.20	(3.31–8.18)		0.885	2.42	(1.40–4.18)	
Absence of ascites	−0.798	0.45	(0.37–0.55)	<0.001	−0.396	0.67	(0.54–0.84)	<0.001
Age (per 5-year increase)	0.153	1.17	(1.12–1.21)	<0.001	0.133	1.14	(1.09–1.19)	<0.001

HR=hazard ratio, CI=confidence interval.

References

- Log rank test tutorial.
- Survival analysis: Self learning book.
- M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman.
  Survival analysis part ii: Multivariate data analysis – an introduction to concepts and methods.
  Nature.
- T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman.
  Survival analysis part i: Basic concepts and first analyses.
  Nature.