Survival Analysis Workshop

Jooyoung Lee

Department of Applied Statistics

Chung-Ang University

Seoul Medical Center

January 06, 2023

Instructor

Jooyoung Lee

- Department of Applied Statistics, Chung-Ang University
- Email: jooylee@cau.ac.kr

Overview

- The objective of this course is to understand statistical methods for survival analysis.
- Application of methods will be presented using the National Health Insurance Sharing Service (NHISS) big data.

Schedule

Week 1: Introduction to Survival Analysis

Week 2: Clinical Study using NHISS I

Week 3: Introduction to Competing Risks Model

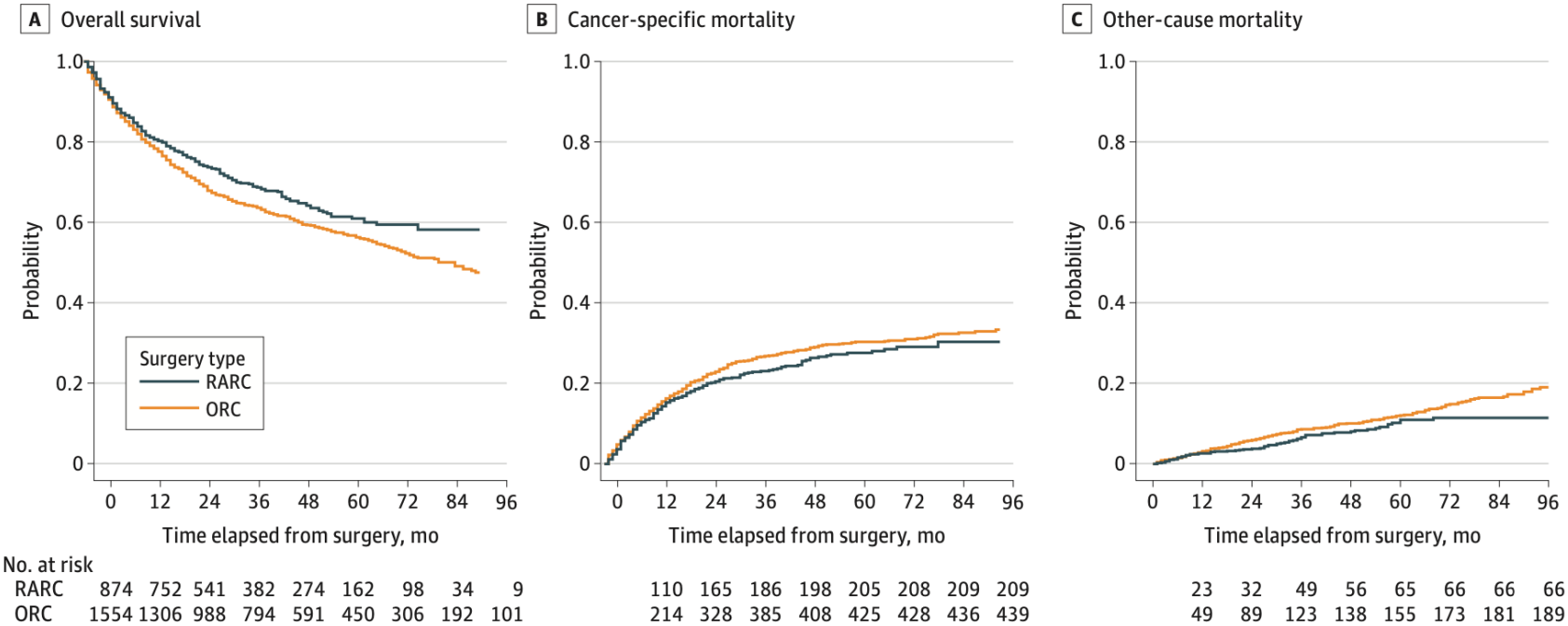
Week 4: Clinical Study using NHISS II

Week 1: Introduction to Survival Analysis

Week 1 Materials

- An introduction to censoring, truncation and examples
- Survival curves, risk set tables, and the Kaplan-Meier estimator
- Log-rank test
- Cox proportional hazard regression models

Example



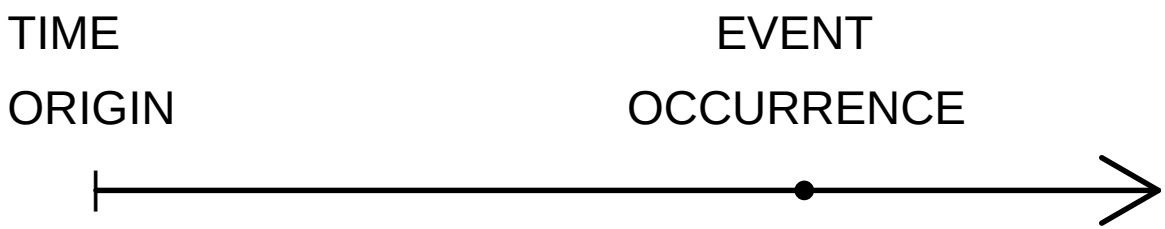
ORC indicates open radical cystectomy; RARC, robot-assisted radical cystectomy.

Figure 1: Figure from Mortezavi, et al., *JAMA Network Open* 2022

Time-to-Event Data

Terminology: Survival time, failure time, lifetime, time-to-event data
Time-to-event data: the time to event measured from some particular starting point.

The definition includes a **time origin**, a **time scale**, and a definition of **the event** of interest.



- Time origin: entry into study, randomization, birth
- Time scale: years, months, weeks, age, mileage
- Event: death, disease occurrence, progression of tumor, hospitalization, recovery

The Purpose of Time-to-Event Data Analysis

- Estimate a event time distribution of a population
- Compare the event time distributions of two or more groups
- Evaluate the effect of risk factors on the event time

Features of Time-to-Event Data

Time-to-event data consist of a mixture of complete and incomplete observations.

Censoring: Censoring arises when the value of a response measurement (event time) is only partially known.

- Right censoring
- Left censoring
- Interval censoring

Truncation: Truncation is termed to describe a condition by which subjects are screened or excluded from the study population.

- Left truncation

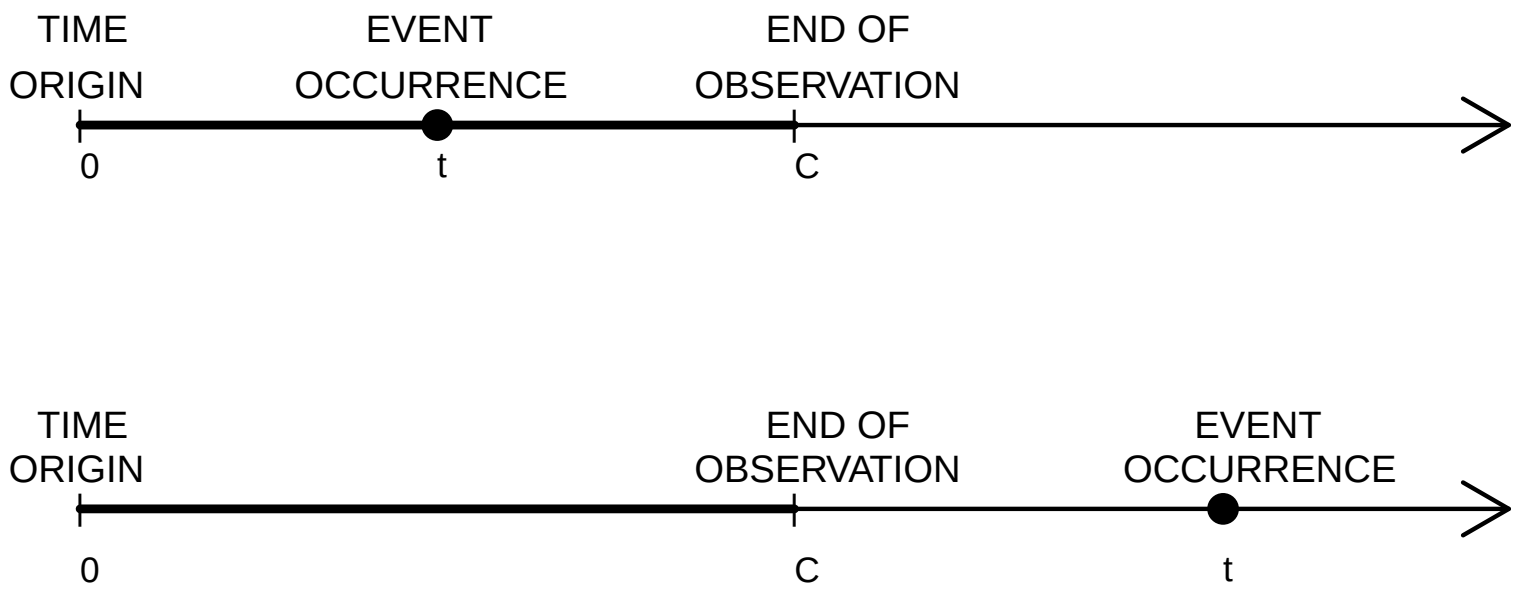
Notation

- T: Event time (survival/failure time) of an individual from a population
- C: Censoring time
- X: Observed time
- δ: Censoring indicator

Right Censored Data

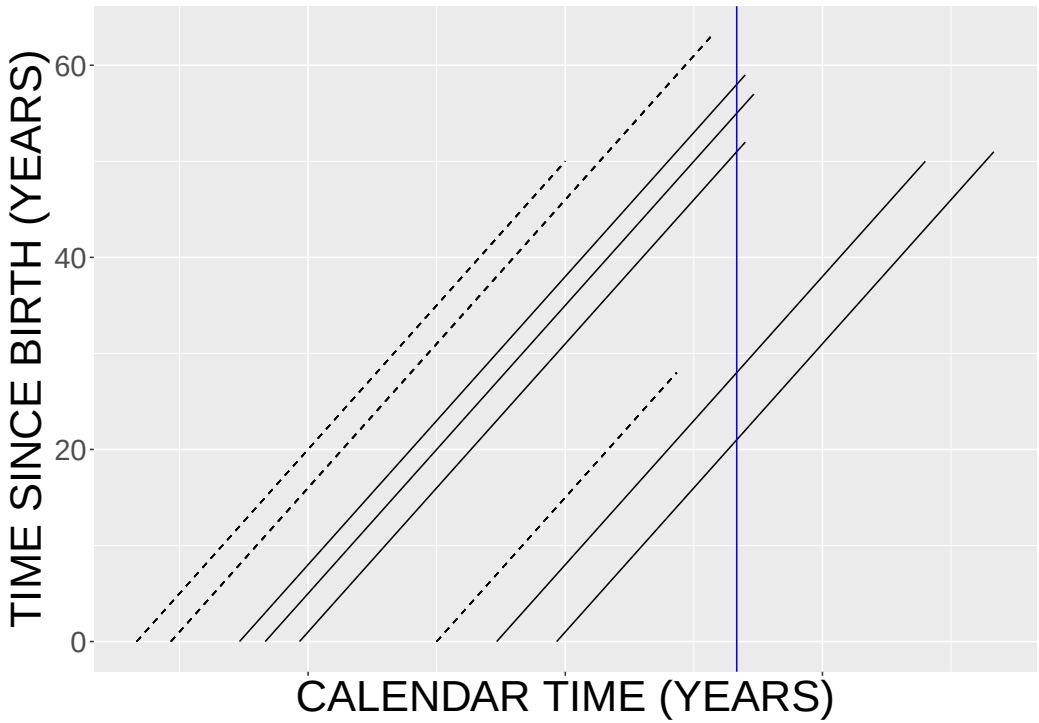
Right censoring: The true unobserved event time is to the right of the censoring time. Right censoring may occur due to no event before the study ends, loss to follow-up or withdrawal from the study.

$$X = \min(T, C), \text{ and } \delta = I(T \leq C)$$



Left Truncated Data

Left truncation arises when there are particular selection conditions for recruiting study participants.

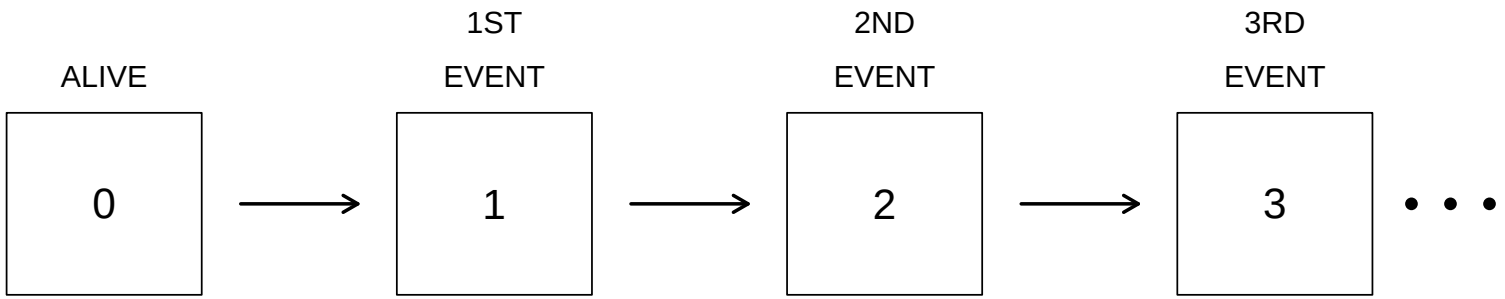


Univariate Survival data


```
library(survival)
head(ovarian)
```

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
6	448	0	56.4301	1	1	2

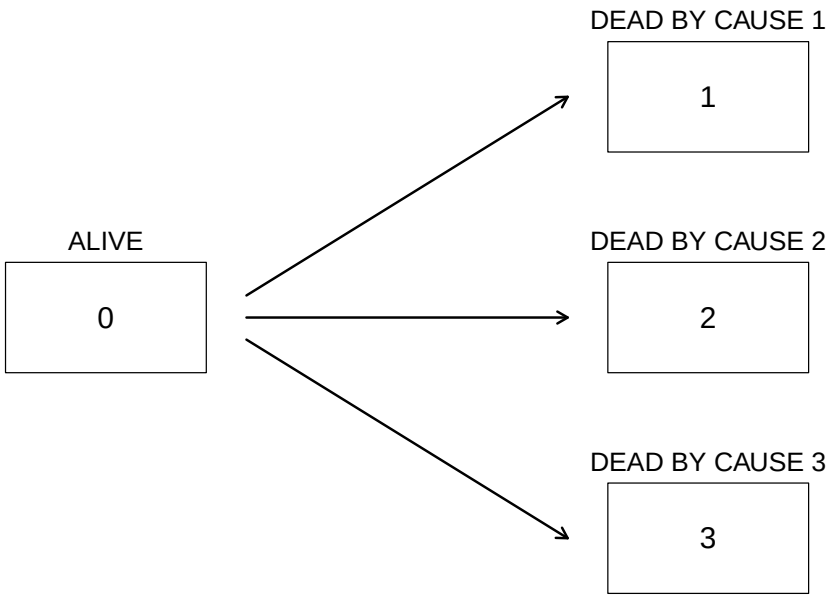
Recurrent Event data



	id	start	stop	status	tstatus	enum	trt
1	1	0	122	1	1	1	1
2	2	0	122	0	1	1	1
3	3	0	3	1	1	1	1
4	3	3	88	1	2	2	1
5	3	88	122	0	2	3	1

Competing Risks data

Competing risks data arise when an individual is at risk of more than one mutually exclusive event.

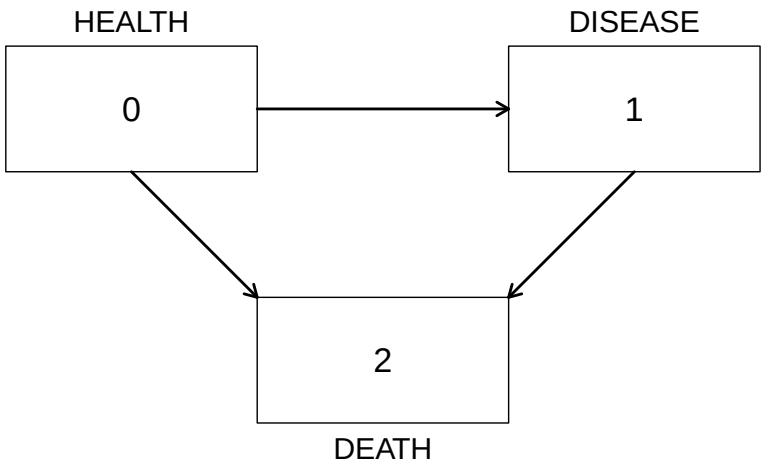



```
library(survival); data(pbc)
pbc[1:5, c(1,2,3,4)]
```

	id	time	status	trt
1	1	400	2	1
2	2	4500	0	1
3	3	1012	2	1
4	4	1925	2	1
5	5	1504	1	2

Multistate Model

Multistate model is a model for a process where subjects move among a finite number of states.




```
library(mstate); data(ebmt3)
ebmt3[1:5, c(1,2,3,4, 5)]
```

	id	prtime	prstat	rfstime	rfsstat
1	1	23	1	744	0
2	2	35	1	360	1
3	3	26	1	135	1
4	4	22	1	995	0
5	5	29	1	422	1

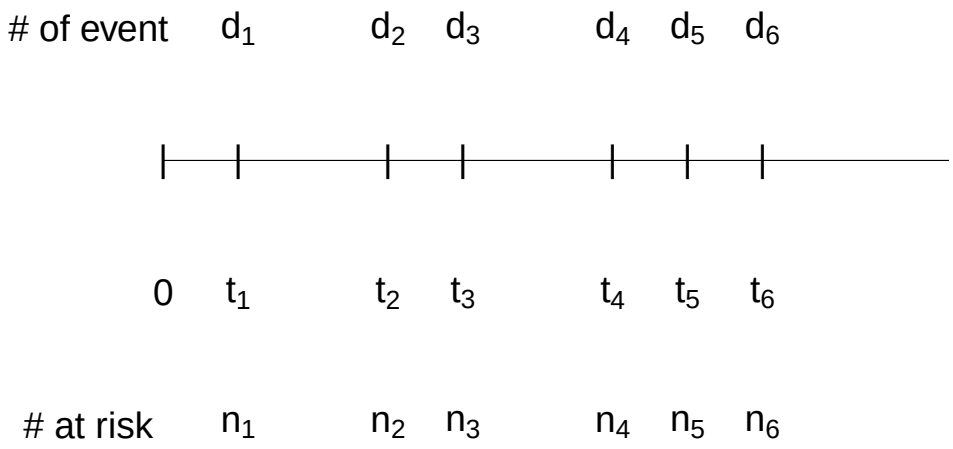
Event time Distribution

Response (T): Time to event (survival/failure time) of an individual from a population.

- The survival function: $S(t) = P(T > t)$.
- The probability density function of T: $f(t) = -dS(t)/dt$.
- The hazard function: $\lambda(t) = \lim_{h \to 0^+} P(t \leq T < t + h | T \geq t)/h$.
- The cumulative hazard function: $\Lambda(t) = \int_0^t \lambda(s)ds$.

Risk Table

- $\{(X_i, \delta_i), i = 1, \ldots, n\}$: a sample of n right-censored survival data.
- $k (k \leq n)$ distinct event times $t_1 < t_2 < \cdots < t_k$.
- n_j: the number of individuals at risk at t_j. The number of individuals with no event and uncensored "just before" time t.
- d_j: the number of individuals who experience the event at t_j.



A risk table displays the number at risk n_j at each time point.

Estimating Survival Curves: the Kaplan-Meier estimator

The Kaplan-Meier estimator $\widehat{S}(t)$ is

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Comparing the Survival of Groups: The Log-Rank Test

We want to test whether the survival curves are the same for two groups.

- $O_j = d_{1j}$: the observed number of events for group 1 at time t_j
- $E_j = n_{1j} d_j / n_j$: the expected number of events for group 1 at time t_j
- $V_j = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$: the variance of the observed number of events for group 1 at time t_j

The logrank test statistic is

$$Z = \frac{\sum_{j=1}^{k} (O_j - E_j)}{\sqrt{\sum_{j=1}^{k} V_j}} \sim N(0, 1) \quad \text{under } H_0$$

Example

Group 0 : 3, 5+, 9, 11+, Group 1 : 8, 12+, 10+, 14+

At $t = 3$

Group 0	1	3	4
Group 1	0	4	4
Total	1	7	8

At $t = 8$

Group 0	0	2	2
Group 1	1	3	4
Total	1	5	6

At $t = 9$

Group 0	1	1	2
Group 1	0	3	3
Total	1	4	5

Example

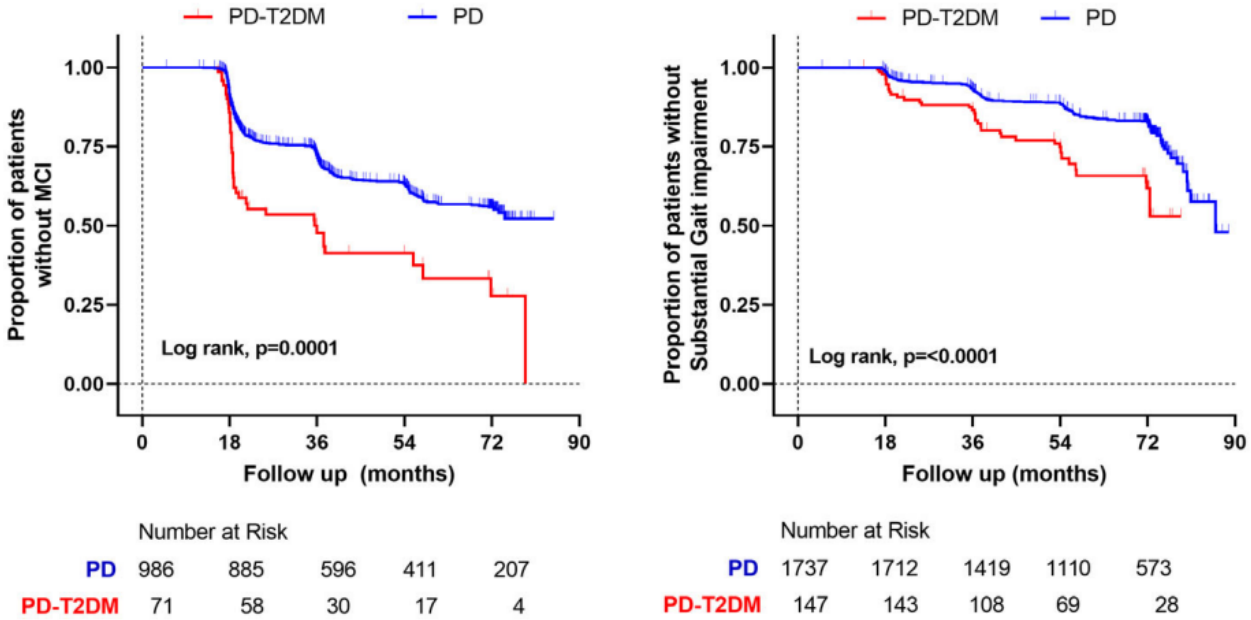


FIG. 3. Timeline for the development of mild cognitive impairment (MCI) and substantial gait impairment, comparing Parkinson's disease (PD) cases with and without type 2 diabetes mellitus (T2DM). Kaplan–Meier curves show the significantly shorter time to develop both of these complications in patients with T2DM. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 2: Figure from Athauda, et al., *Movement Disorder* 2022

Cox Proportional Hazard Regression Model

The Cox (1972) proposed the proportional hzard regression model:

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\mathbf{z}'\beta)$$

- A vector of covariates $\mathbf{z} = (z_1, \ldots, z_p)^T$
- A baseline hazard function $\lambda_0(t)$.

Hazard Ratio

- The estimated hazard ratio (HR) $= \exp(\hat{\beta}_j)$
- A 95% CI for HR $= \exp(\hat{\beta}_j \pm 1.96 s.e.(\hat{\beta}_j))$
- p-value $= P(|Z| > |z_0|)$, where $z_0 = \hat{\beta}_j / s.e.(\hat{\beta}_j)$, $Z \sim N(0,1)$ under $H_0 : \beta_j = 0$.

Stratified Cox Regression Models

- Often times we are interested in assessing the effects of exposures while adjusting for other factors in a regression analysis.
- We can stratify on the factor we want to control for if we are not interested in the hazard ratio for this factor to avoid the proportional hazard assumption.

A stratified Cox regression model is formed by

$$\lambda_k(t|\mathbf{z}_k) = \lambda_{k0}(t)\exp(\mathbf{z}_k'\beta),$$

where $\lambda_{k0}(t)$ is the baseline hazard for stratum k and β represents common covariate effects.

Extended Cox Models: Time-dependent Covariates

Consider the study for transplant on survival.

- The Cox model would compare the survival distributions between those without a transplant (ever) to those with a transplant.
- A transplant status was determined at the end of the study and this value characterizes a subject for the entire follow-up period.
- Patients who died early had less time available to get transplants.
- Therefore, the survival estimates are overestimated for the patients who had transplants, and we call this problem as "immortal bias".

Extended Cox Models: Time-dependent Covariates

- As the status of transplant changes over time, we consider a model with a time-dependent indicator of whether a patient had a transplant at each point.
- We compare the risk of an event between the patients who had already transplants and who had not yet transplants at each event time.
- We recalculate the number at risk for each group at each event time.

A path of the covariate upto t: $\bar{\mathbf{z}}_i(t) = \{\mathbf{z}_i(u), 0 \leq u \leq t\}$.

We model

$$\lambda(t|\bar{\mathbf{z}}(t)) = \lambda_0(t)\exp(\mathbf{z}(t)'\beta).$$

Reference

- Mortezavi, A., Crippa, A., Kotopouli, M. I., Akre, O., Wiklund, P., & Hosseini, A. (2022). Association of Open vs Robot-Assisted Radical Cystectomy With Mortality and Perioperative Outcomes Among Patients With Bladder Cancer in Sweden. JAMA network open, 5(4), e228959-e228959.
- Athauda, D., Evans, J., Wernick, A., Virdi, G., Choi, M. L., Lawton, M., ... & Gandhi, S. (2022). The impact of type 2 diabetes in Parkinson's disease. Movement Disorders, 37(8), 1612-1623.