

# Statistical Practice in Epidemiology 2018

Survival analysis with competing risks

Janne Pitkaniemi (EL)

# Points to be covered

1. Survival or time to event data & censoring.
2. Competing risks: event-specific cumulative incidences & hazards.
3. Kaplan–Meier and Aalen–Johansen estimators.
4. Regression modelling of hazards: Cox model.
5. Packages `survival`, `mstate`, `cmprisk`.
6. Functions `Surv()`, `survfit()`, `plot.survfit()`, `coxph()`, `Cuminc()`.

# Survival time – time to event

**Time** spent ( $\text{lex.dur}$ ) in a given **state** ( $\text{lex.Cst}$ ) from its beginning till a certain *endpoint* or *outcome event* ( $\text{lex.Xst}$ ) or *transition* occurs, changing the state to another.

Examples of such times and outcome events:

- ▶ lifetime: birth  $\rightarrow$  death,
- ▶ duration of marriage: wedding  $\rightarrow$  divorce,
- ▶ healthy exposure time:  
start of exposure  $\rightarrow$  onset of disease,
- ▶ clinical survival time:  
diagnosis of a disease  $\rightarrow$  death.

## Ex. Survival of 338 oral cancer patients

Important variables:

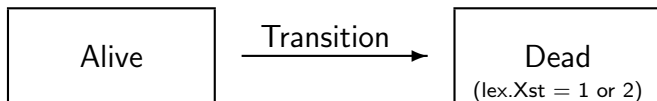
- ▶ `time` = duration of patientship from diagnosis (**entry**) till death (`death`) or censoring (`Alive`), (`lex.Cst` is (`Alive`))
- ▶ `event` = indicator for the outcome and its observation at the end of follow-up (**exit**):  
0 = censoring,  
1 = death from oral cancer

Special features:

- ▶ Two possible endpoints
- ▶ Censoring – incomplete observation of the survival time.

# Set-up of classical survival analysis

- ▶ **Two-state model:** only one type of event changes the initial state.
- ▶ Major applications: analysis of lifetimes since birth and of survival times since diagnosis of a disease until death from any cause.



- ▶ **Censoring:** Death and final lifetime not observed for some subjects due to emigration or closing the follow-up while they are still alive

# Distribution concepts: hazard function

The **hazard rate** or **intensity** function  $\lambda(t)$

$$\lambda(t) = P(t < T \leq t + \Delta | T > t) / \Delta, \text{ for small } \Delta$$

$\approx$  the conditional probability that the event occurs in a short interval  $(t, t + \Delta]$ , given that it does not occur before  $t$ , divided by interval length.

In other words, during a short interval

$$\text{risk of event} \approx \text{hazard} \times \text{interval length}$$

# Distribution concepts: survival and cumulative hazard functions

## Survival function

$$S(t) = P(T > t),$$

= probability of avoiding the event at least up to  $t$   
(the event occurs only after  $t$ ).

The **cumulative hazard** (or integrated intensity):

$$\Lambda(t) = \int_0^t \lambda(u) du$$

Connections between the functions:

$$S(t) = \exp\{-\Lambda(t)\}$$

# Observed data on survival times

For individuals  $i = 1, \dots, n$  let

$T_i$  = time to outcome event,

$U_i$  = time to censoring.

Censoring is assumed **noninformative**, *i.e.*  
independent from occurrence of events.

We observe

$y_i = \min\{T_i, U_i\}$ , *i.e.* the exit time, and

$\delta_i = 1_{\{T_i < U_i\}}$ , indicator (1/0) for the outcome event  
occurring first, before censoring.

Censoring must properly be taken into account in the  
statistical analysis.



# Approaches for analysing survival time

- ▶ **Parametric model** (like Weibull, gamma, etc.) on hazard rate  $\lambda(t) \rightarrow$  Likelihood:

$$L = \prod_{i=1}^n \lambda(y_i)^{\delta_i} S(y_i)$$

- ▶ **Piecewise constant rate** model on  $\lambda(t)$ 
  - see Bendix's lecture on time-splitting (Poisson likelihood).
- ▶ **Non-parametric** methods, like Kaplan–Meier (KM) estimator of survival curve  $S(t)$  and Cox proportional hazards model on  $\lambda(t)$ .

# R package survival

Tools for analysis with one outcome event.

- ▶ `Surv(time, event) -> sobj`  
creates a **survival object** `sobj` assuming that all start at 0, containing pairs  $(y_i, \delta_i)$ ,
- ▶ `Surv(entry, exit, event) -> sobj2`  
creates a survival object from entry and exit times,
- ▶ `survfit(sobj ~ x) -> sfo`  
creates a **survfit** object `sfo` containing KM or other non-parametric estimates (also from a fitted Cox model),
- ▶ `plot(sfo)`  
plot method for survival curves and related graphs,
- ▶ `coxph(sobj ~ x1 + x2)`  
fits a Cox model with covariates `x1` and `x2`.
- ▶ `survreg()` – parametric survival models.

## Ex. Oral cancer data (cont'd)

```
> orca$suob <- Surv(orca$time, 1*(orca$event > 0) )

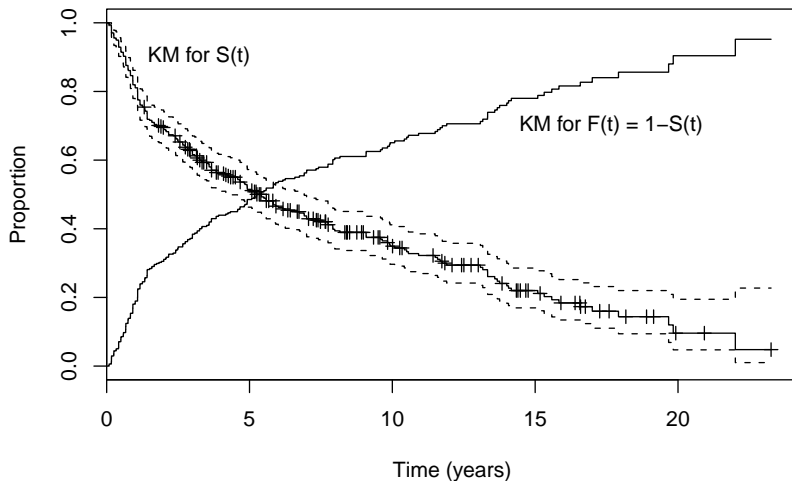
> orca$suob[1:7]    # + indicates censored observation
[1] 5.081+ 0.419  7.915  2.480  2.500  0.167  5.925+

> km1 <- survfit( suob ~ 1, data = orca)
> km1               # brief summary
records    n.max n.start  events  median 0.95LCL 0.95UCL
 338.00   338.00  338.00  229.00   5.42    4.33    6.92

> summary(km1)      # detailed KM-estimate
   time n.risk n.event survival std.err lower 95\% CI upper 95\% CI
0.085   338     2  0.9941 0.00417   0.9859    1.000
0.162   336     2  0.9882 0.00588   0.9767    1.000
0.167   334     4  0.9763 0.00827   0.9603    0.993
0.170   330     2  0.9704 0.00922   0.9525    0.989
0.246   328     1  0.9675 0.00965   0.9487    0.987
...
```

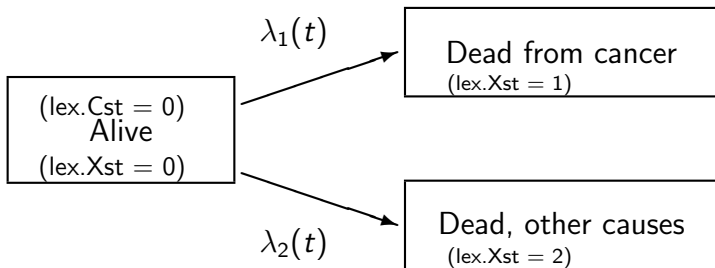
# Oral cancer: Kaplan-Meier estimates

**Estimated survival (+censorings & conf.limits) and CDF**



## Competing risks model: causes of death

- ▶ Often the interest is focused on the risk or hazard of dying from one specific cause.
- ▶ That cause may eventually not be realized, because a **competing cause** of death hits first.



- ▶ Generalizes to several competing causes.

# Competing events & competing risks

In many epidemiological and clinical contexts there are competing events that may occur before the target event and remove the person from the population at risk for the event, *e.g.*

- ▶ *target event*: occurrence of endometrial cancer,  
*competing events*: hysterectomy or death.
- ▶ *target event*: relapse of a disease  
(ending the state of remission),  
*competing event*: death while still in remission.
- ▶ *target event*: divorce,  
*competing event*: death of either spouse.

## Event-specific quantities

**Cumulative incidence function (CIF)** or

$$F_c(t) = P(T \leq t \text{ and } C = c), \quad c = 1, 2,$$

From these one can recover

- ▶  $F(t) = \sum_c F_c(t)$ , CDF of event-free survival time  $T$ , *i.e.* cumulative risk of any event by  $t$ .
- ▶  $S(t) = 1 - F(t)$ , **event-free survival function**, *i.e.* probability of avoiding all events by  $t$ , but  $S(t) \neq F_1(t) + F_2(t)$

## Event-specific quantities (cont'd)

### Event- or cause-specific hazard function

$$\begin{aligned}\lambda_c(t) &= \lim_{\Delta \rightarrow 0} \frac{P(t < T \leq t + \Delta \text{ and } C = c \mid T > t)}{\Delta} \\ &= \frac{f_c(t)}{1 - F(t)}\end{aligned}$$

CIF = risk of event  $c$  over risk period  $[0, t]$  in the presence of competing risks, also obtained

$$F_c(t) = \int_0^t \lambda_c(v) S(v) dv, \quad c = 1, 2,$$



# Warning of “net risk” and “cause-specific survival”

- ▶ The “**net risk**” of outcome  $c$  by time  $t$ , assuming hypothetical elimination of competing risks, is often defined as

$$F_1^*(t) = 1 - S_1^*(t) = 1 - \exp\{-\Lambda_1(t)\} \neq S(t)$$

- ▶ In clinical survival studies, function  $S_1^*(t)$  is often called “**cause-specific survival**”, or “**net survival**”
- ▶ Yet, these \*-functions,  $F_1^*(t)$  and  $S_1^*(t)$ , lack proper probability interpretation when competing risks exist.
- ▶ Hence, their use should be viewed critically (Andersen & Keiding, *Stat Med*, 2012)

## Example: Risk of lung cancer by age $a$ ?

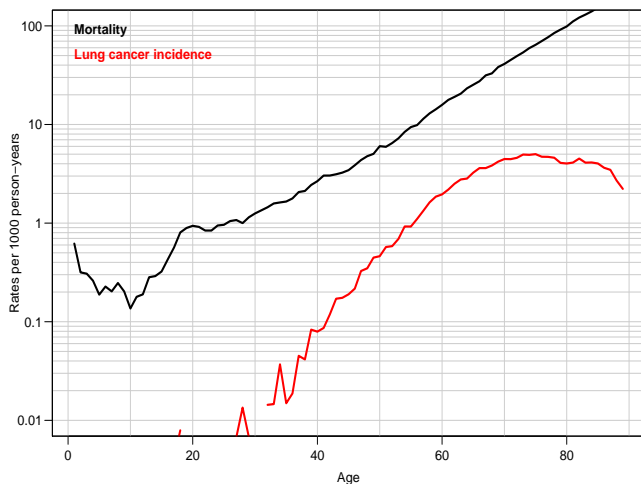
- ▶ Nordcan & Globocan give “**cumulative risk**” by 75 y of age, computed from  $1 - \exp\{-\text{CR}(75)\}$ , as an estimate of the probability of getting cancer before age 75 y, assuming that death were avoided by that age. This is based on deriving “net risk” from cumulative hazard:

$$F_1^*(a) = 1 - \exp\{-\Lambda_c(a)\}.$$

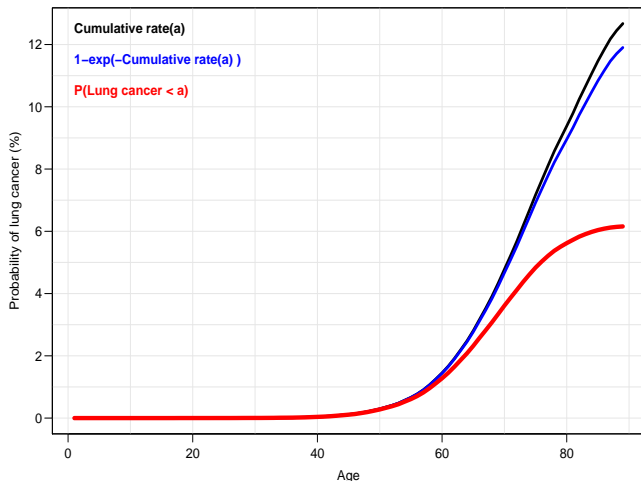
- ▶ Yet, cancer occurs in a mortal population.
- ▶ As such  $\text{CR}(75)$  is a sound age-standardized summary measure for comparing cancer incidence across populations based on a neutral standard population.

## Example. Male lung cancer in Denmark

Event-specific hazards:  $\lambda_1(a)$  (lung cancer) death  $\lambda_2(a)$  by age estimated by age-spec. rates.



# Cumulative incidence of lung cancer by age



Both CR and  $1 - \exp(-\text{CR})$  tend to overestimate the real cumulative incidence CI after 60 y.

# Analysis with competing events

Let  $U_i$  = censoring time,  $T_i$  = time to first event, and  $C_i$  = variable for event 1 or 2. We observe

- ▶  $y_i = \min\{T_i, U_i\}$ , i.e. the exit time, and
- ▶  $\delta_{ic} = 1_{\{T_i < U_i \text{ \& } C_i = c\}}$ , indicator (1/0) for event  $c$  being first observed,  $c = 1, 2$ .

Non-parametric estimation of CIF

- ▶ Let  $t_1 < t_2 < \dots < t_K$  be the  $K$  distinct time points at which any outcome event was observed,  
Let also  $\hat{S}(t)$  be KM estimator for overall  $S(t)$ .
- ▶ **Aalen-Johansen estimator** (AJ) for the cumulative incidence function  $F(t)$  should be used

# R tools for competing risks analysis

## Package mstate

- ▶ `Cuminc(time, status, ...)`:  
AJ-estimates (and SEs) for each event type (status, value 0 indicating censoring)

## Package cmprsk

- ▶ `cuminc(ftime, fstatus, ...)` computes CIF-estimates, `plot.cuminc()` plots them.
- ▶ `crr()` fits Fine–Gray models for the hazard  $\gamma_c(t)$  of the subdistribution

## Package Epi – Lexis tools for multistate analyses

- ▶ will be advertised by Bendix!

## Ex. Survival from oral cancer

- ▶ Creating a Lexis object with two outcome events and obtaining a summary of transitions

```
> orca.lex <- Lexis(exit = list(stime = time),  
                    exit.status = factor(event,  
                    labels = c("Alive", "Oral ca. death", "Other death") ),  
                    data = orca)
```

```
> summary(orca.lex)
```

Transitions:

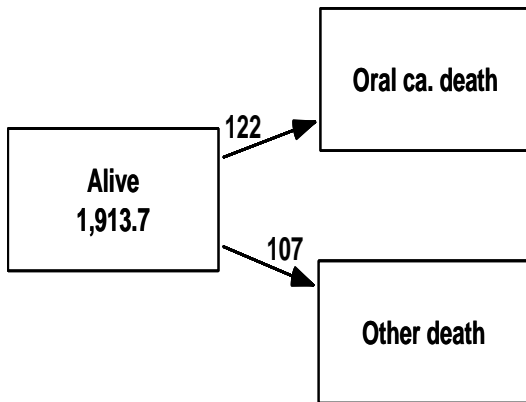
To

From	Alive	Oral	ca.	Other	Records:	Events:	Risk time:
Alive	109		122	107	338	229	1913.67

# Box diagram for transitions

Interactive use of function boxes().

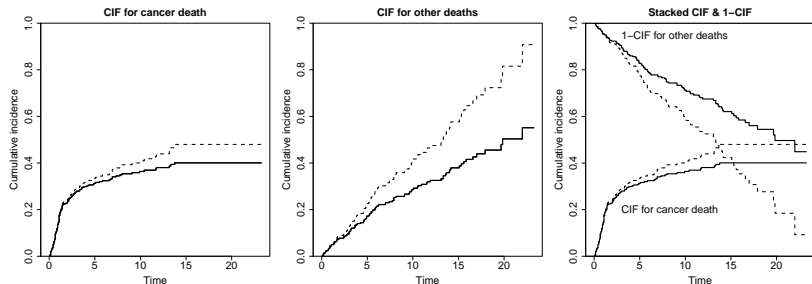
```
> boxes(orca.lex)
```





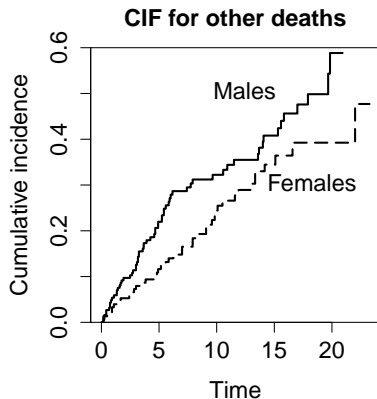
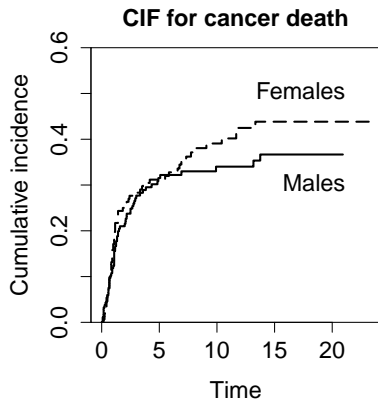
## Ex. Survival from oral cancer

- ▶ AJ-estimates of CIFs (solid) for both causes.
- ▶ Naive KM-estimates of CIF (dashed)  $>$  AJ-estimates
- ▶ CIF curves may also be stacked (right).



**NB.** The sum of the naive KM-estimates of CIF exceeds 100% at 13 years!

## Ex. CIFs by cause in men and women



CIF for cancer higher in women (chance?) but for other causes higher in men (no surprise).