Epidemiologic Data Analysis using R
Part 8:
Competing risk in time-to-event analysis

Janne Pitkäniemi

Finnish Cancer Registry, Finland, <janne.pitkaniemi@cancer.fi>

University of Tampere
Faculty of Social Sciences
Feb 26- Apr 9 2018

Points to be covered

1. Distribution concepts for times to event:
   survival, hazard and cumulative hazard,
2. Competing risks: event-specific cumulative incidences & hazards.
4. Kaplan–Meier and Aalen–Johansen estimators.
5. Regression modelling of competing hazards:
6. Packages `survival`, `mstate`, `cmprisk`.
7. Functions `Surv()`, `survfit()`, `plot.survfit()`, `coxph()`, `Cuminc()`.

Points not to be covered – many!

Survival time – time to event

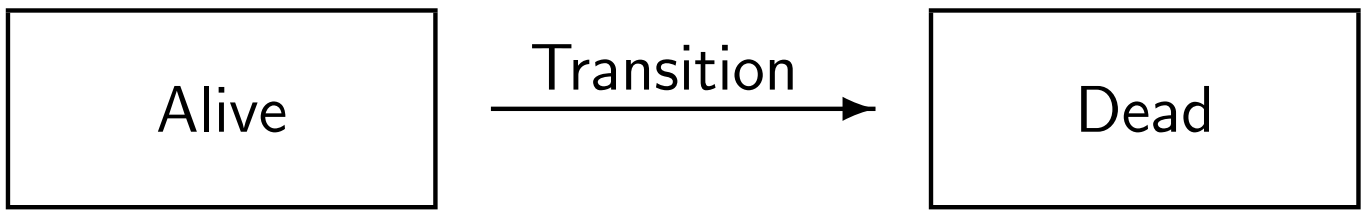
Let T be the **time** spent in a given **state** from its beginning till a certain *endpoint* or *outcome* **event** or *transition* occurs, changing the state to another.
(`lex.Cst - lex.dur - lex.Xst`)

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.

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: survival function

Cumulative distribution function (CDF) $F(t)$ and density function $f(t) = F'(t)$ of survival time T:

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

$=$ **risk** or probability that the event occurs by t.

Survival function

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du,$$

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

Distribution concepts: hazard and cumulative function

The **hazard rate** or **intensity** function $h(t)$

$$\lambda(t) = \lim_{\Delta \to 0} P(t < T \leq t + \Delta | T > t)/\Delta$$
$$= \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta)}{P(T > t)} \frac{1}{\Delta} = \frac{f(t)}{S(t)}$$

In other words, during a short interval

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

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

$$\Lambda(t) = \int_0^t \lambda(v) dv$$

Approaches for analysing survival time

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

Ex. Survival of 338 oral cancer patients

Important variables:

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

Special features:

- Several possible endpoints, *i.e.* alternative causes of death, of which only one is realized.
- Censoring – incomplete observation of the survival time.

Ex. Oral cancer data

```
> head(orca)
     sex      age stage  time event
1   Male 65.42274  unkn 5.081     0
2 Female 83.08783   III 0.419     1
3   Male 52.59008    II 7.915     2
4   Male 77.08630     I 2.480     2
5   Male 80.33622    IV 2.500     1
6 Female 82.58132    IV 0.167     2
```

Ex. Oral cancer data

Analysis of overall motality

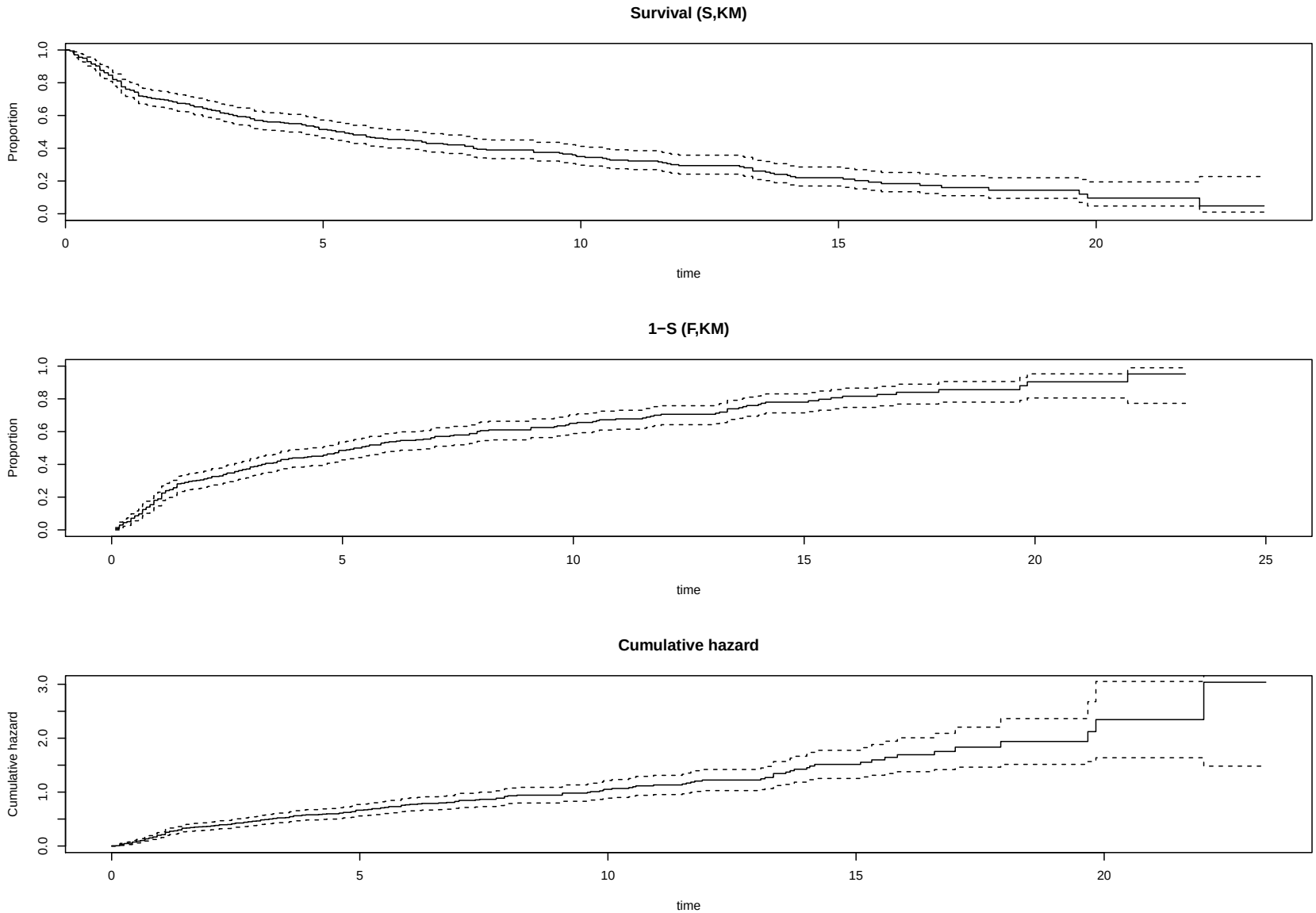
```
> orca$suob <- Surv(orca$time, 1*(orca$event > 0) )
```

```
> 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
```

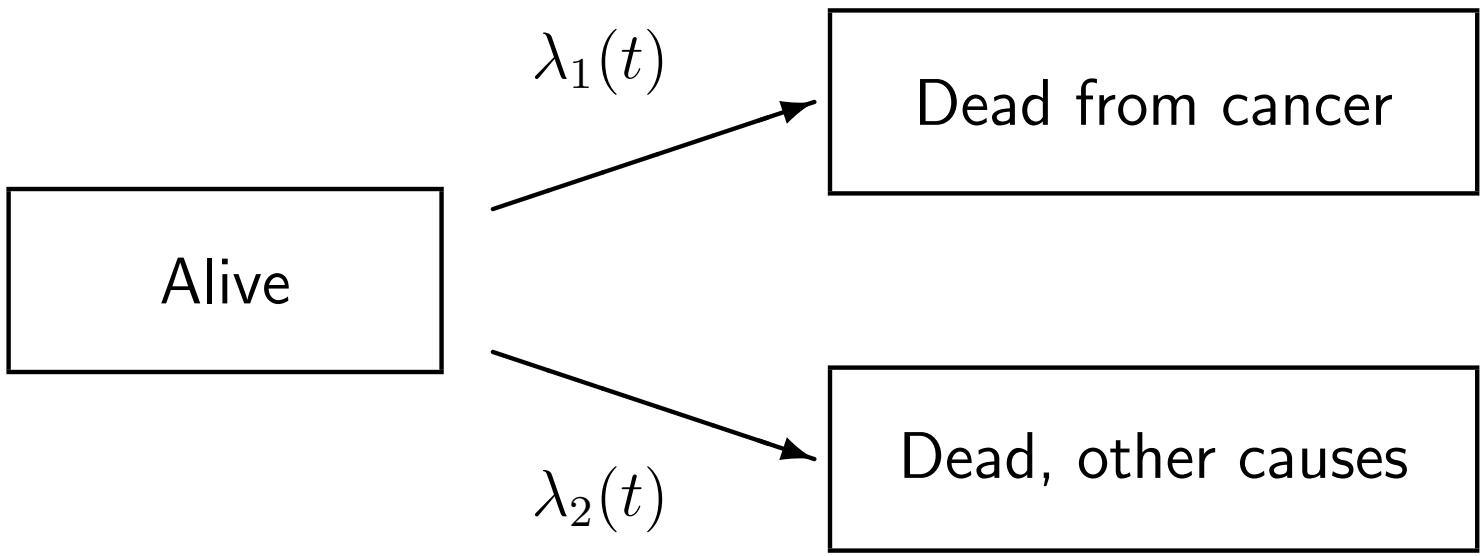
```
   time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.085    338       2   0.9941 0.00417       0.9859        1.000
  0.162    336       2   0.9882 0.00588       0.9767        1.000
  0.167    334       4   0.9763 0.00827       0.9603        0.993
  0.170    330       2   0.9704 0.00922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

Oral cancer: Kaplan-Meier estimates



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

- **a competing risk** is an event that either hinders the observation of the event of interest or modifies the chance that this event occurs.
- In those cases, the competing risk hinders the observation of the event of interest or modifies the chance that this event occurs.
- For example, when performing a study with mortality on dialysis as the outcome of interest, a patient may receive a kidney transplant. This transplant is a competing risk because after the transplantation, this patient is not on dialysis anymore and therefore no longer at risk of dying while being on dialysis. In this case, the competing event, i.e. receiving a kidney transplant, hinders the occurrence of the event of interest.

Research question is relevant for the method

- There are two main methodological strems: **hazard and subdistribution approach** to estimation of parameters of interest.
- What kind of research question one aims to answer. In general, there are two types of research questions which can be answered with epidemiological studies.
- Aetiological research aims to investigate the causal relationship between risk factors or determinants and a given outcome. The effect is measures in HR.
- Prognostic research aims to predict the probability of a given outcome at a given time for an individual patient.

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 **subdistribution function** for event c:

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

subdensity function $f_c(t) = dF_c(t)/dt$

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

Event-specific quantities (cont'd)

Event- or **cause-specific hazard function**

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

$\approx$ Risk of *event* c in a short interval $(t, t + \Delta]$, given *avoidance of all events* up to t, per interval length.

Event- or **cause-specific cumulative hazard**

$$\Lambda_c(t) = \int_0^t \lambda_c(v) dv$$

Event-specific quantities (cont'd)

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

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

- Depends on the hazard of the competing event, too, via

$$\begin{aligned} S(t) &= \exp\left\{ - \int_0^t [\lambda_1(v) + \lambda_2(v)] dv \right\} \\ &= \exp\{-\Lambda_1(t)\} \times \exp\{-\Lambda_2(t)\}. \end{aligned}$$

Hazard of the subdistribution

$$\gamma_c(t) = f_c(t)/[1 - F_c(t)]$$

- Is not the same as $\lambda_c(t) = f_c(t)/[1 - F(t)]$,
- Interpretation tricky!

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

$$F_c^*(t) = 1 - S_c^*(t) = 1 - \exp\{-\Lambda_c(t)\}$$

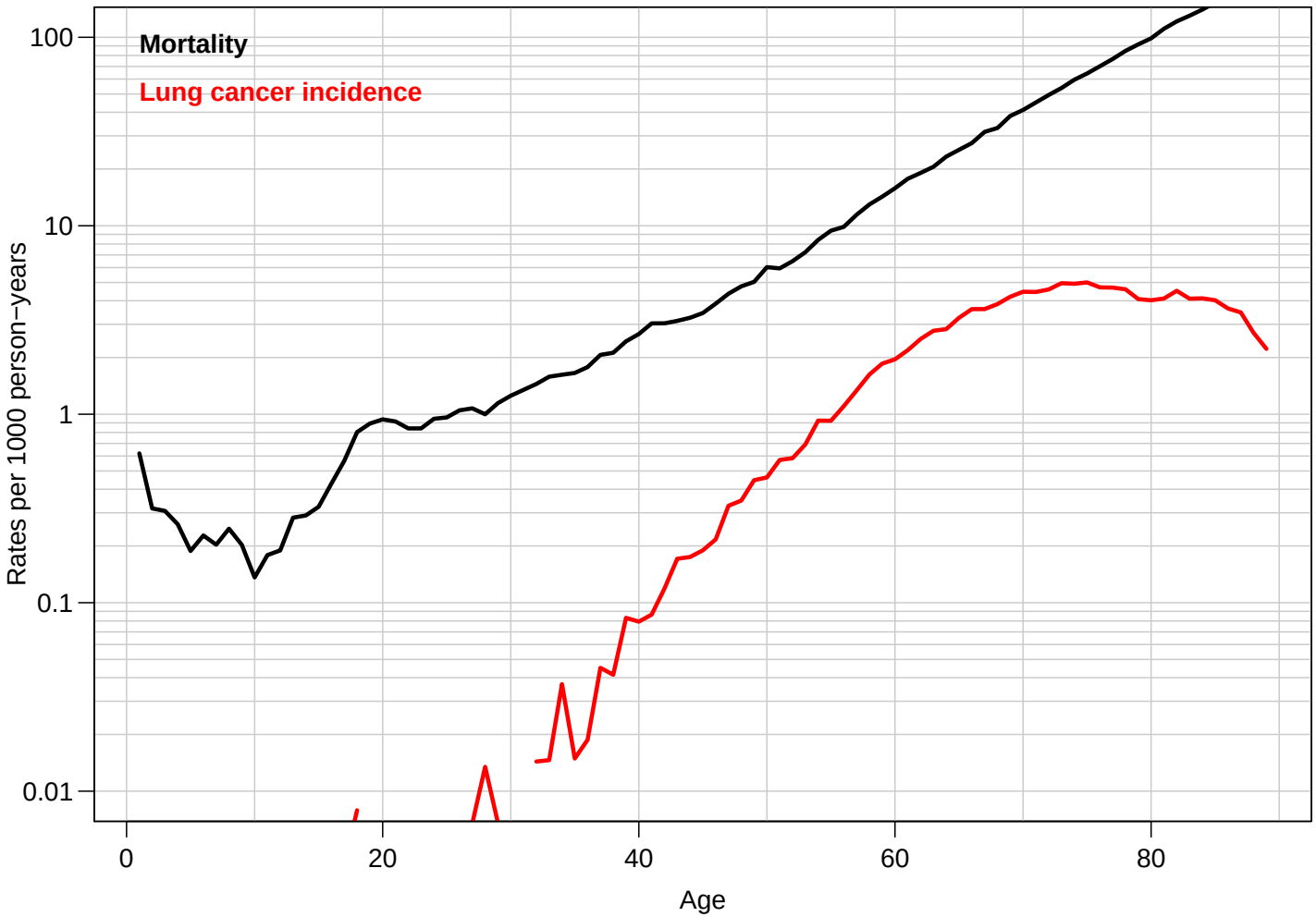
- In clinical survival studies, function $S_c^*(t)$ is often called "**cause-specific survival**", and estimated by KM, but treating competing deaths as censorings.
- Yet, these *-functions, $F_c^*(t)$ and $S_c^*(t)$, lack proper probability interpretation when competing risks exist.
- Hence, their use and naive KM estimation should be viewed critically (Andersen & Keiding, *Stat Med*, 2012)

Example: Risk of lung cancer by age a?

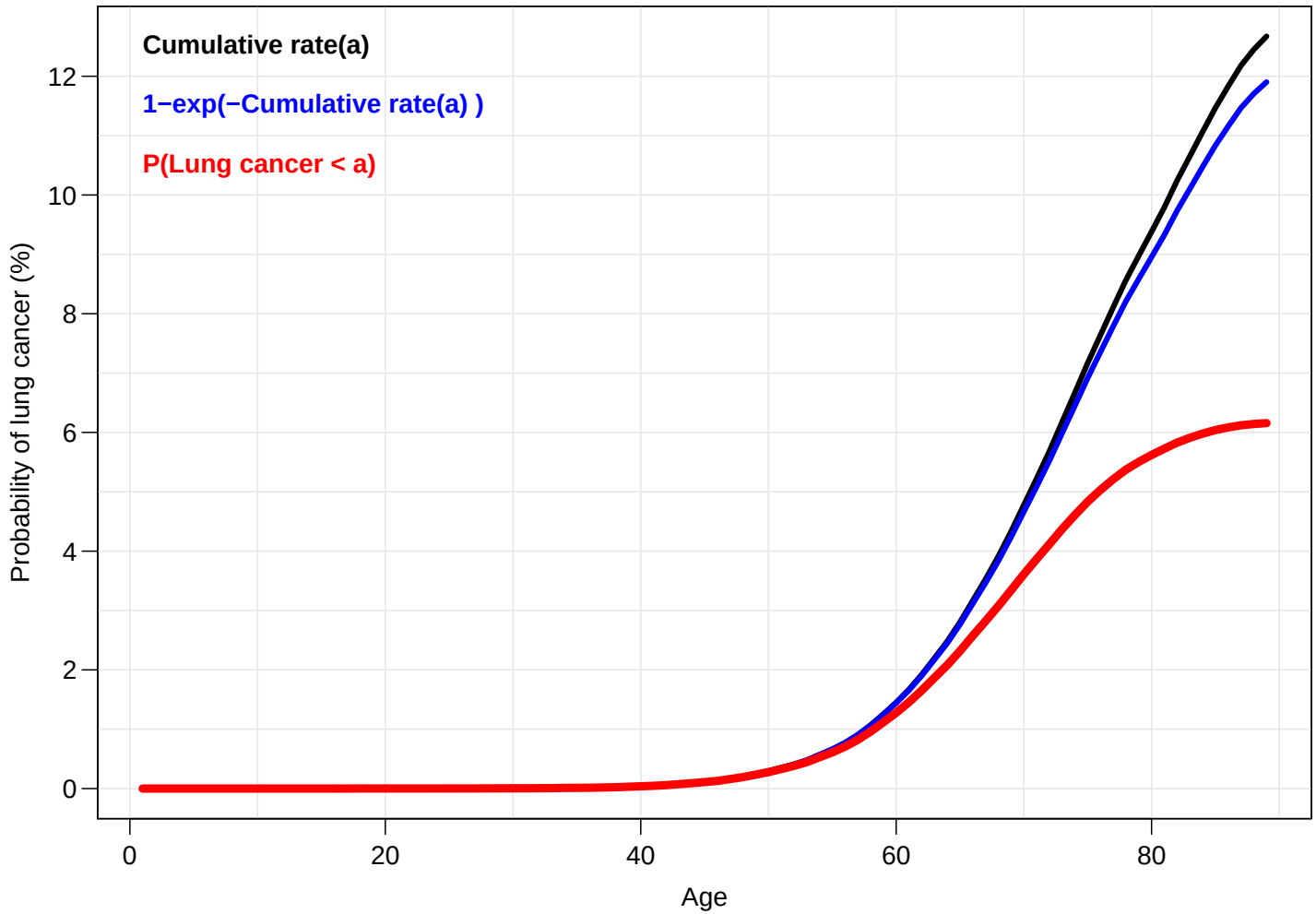
- Empirical **cumulative rate** $\mathrm{CR}(a) = \sum_{k<a} I_k \Delta_k$, i.e. ageband-width (Δ_k) weighted sum of empirical age-specific incidence rates I_k up to a given age a $=$ estimate of cumulative hazard $\Lambda_c(a)$.
- Nordcan & Globocan give "**cumulative risk**" by 75 y of age, computed from $1 - \exp\{-\mathrm{CR}(75)\}$, 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:
$$F_c^*(a) = 1 - \exp\{-\Lambda_c(a)\}.$$
- Yet, cancer occurs in a mortal population.
- As such $\mathrm{CR}(75)$ 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 $\lambda_c(a)$ by age estimated by age-spec. rates of death and lung ca., resp.



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.

Non-parametric estimation of CIF

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

$$\widetilde{F}_c(t) = \sum_{t_k \le t} \frac{D_{kc}}{n_k} \times \widetilde{S}(t_{k-1}), \quad \text{where}$$

  n_k = size of the risk set at t_k $(k = 1, \ldots, K)$,
  D_{kc} = no. of cases of event c observed at t_k.
- Naive KM estimator $\widetilde{F}^*_c(t)$ of "net survival" treats competing events occuring first as censorings:

$$\widetilde{F}^*_c(t) = 1 - \widetilde{S}^*_c(t) = 1 - \prod_{t_k \le t} \frac{n_k - D_{kc}}{n_k}$$

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

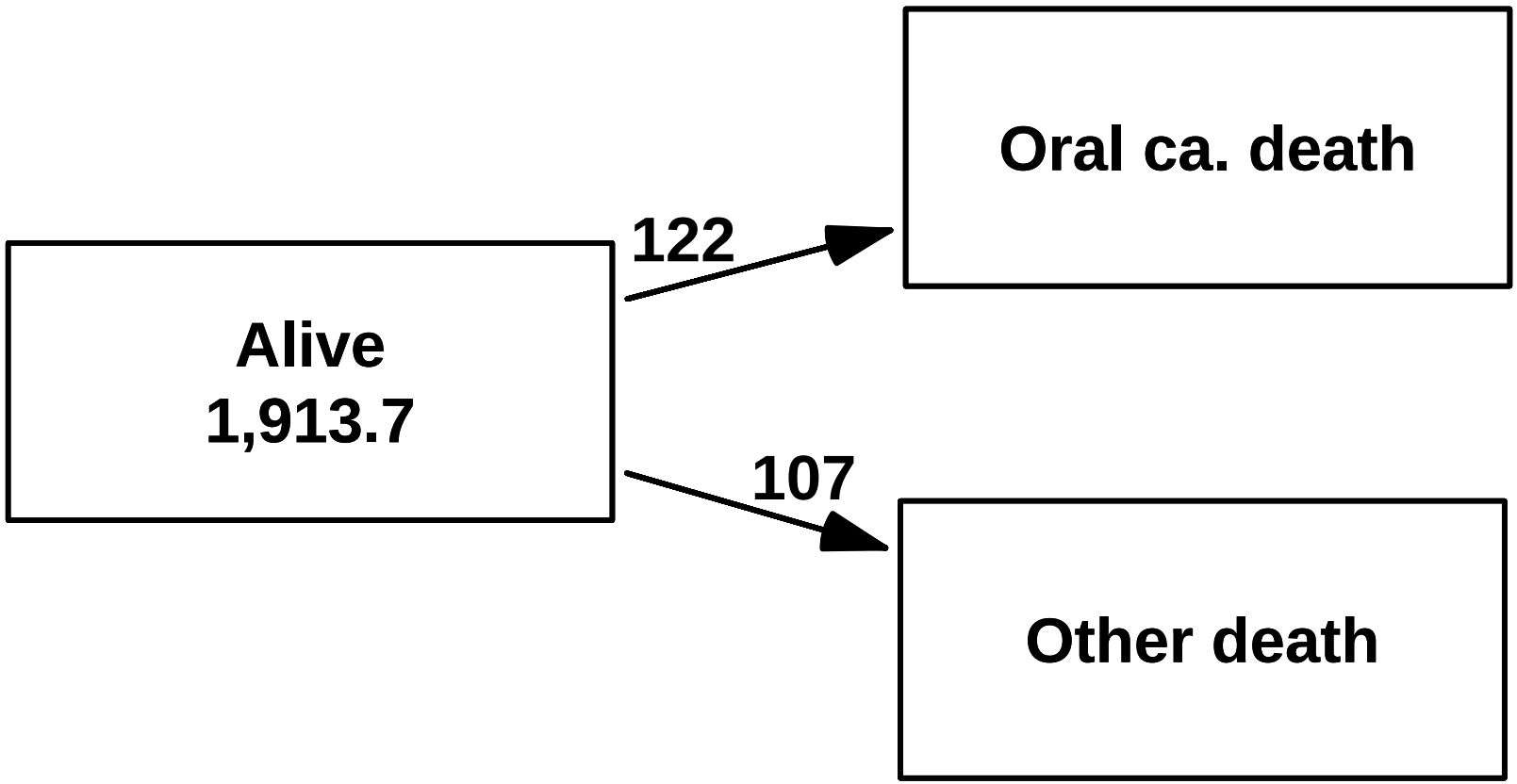
```
> orca.lex <- Lexis(exit = list(stime = time),
                    exit.status = factor(event,
      labels = c("Alive", "Oral ca. death", "Other death") ),
                         data = orca)

> 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

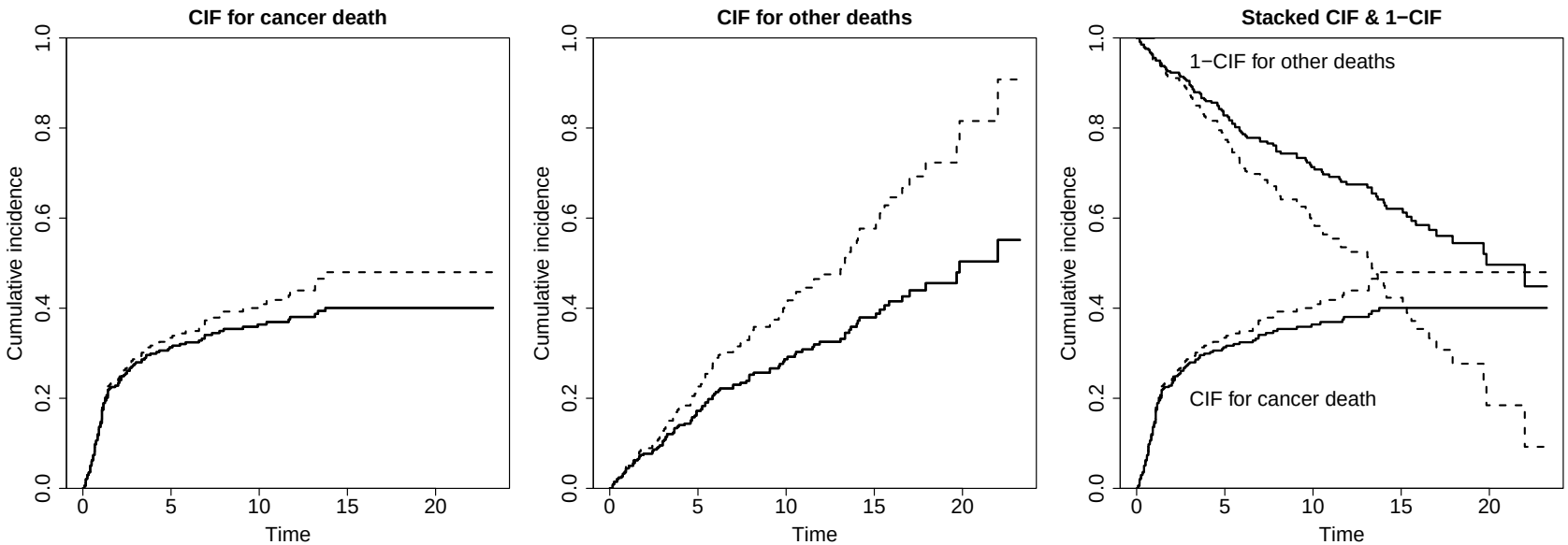
Interactive use of function `boxes()`.

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



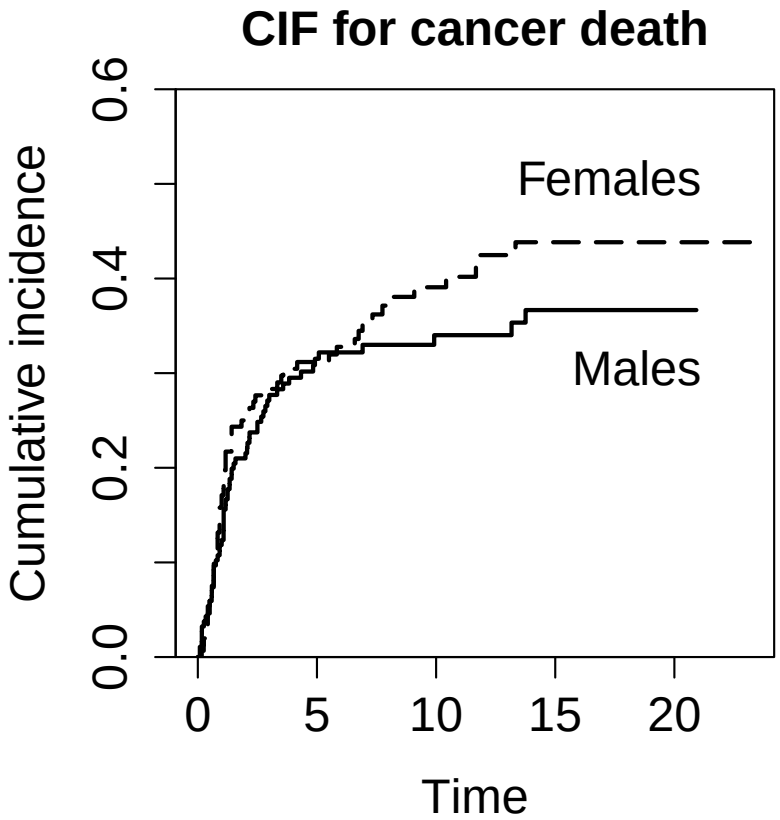
Ex. Survival from oral cancer

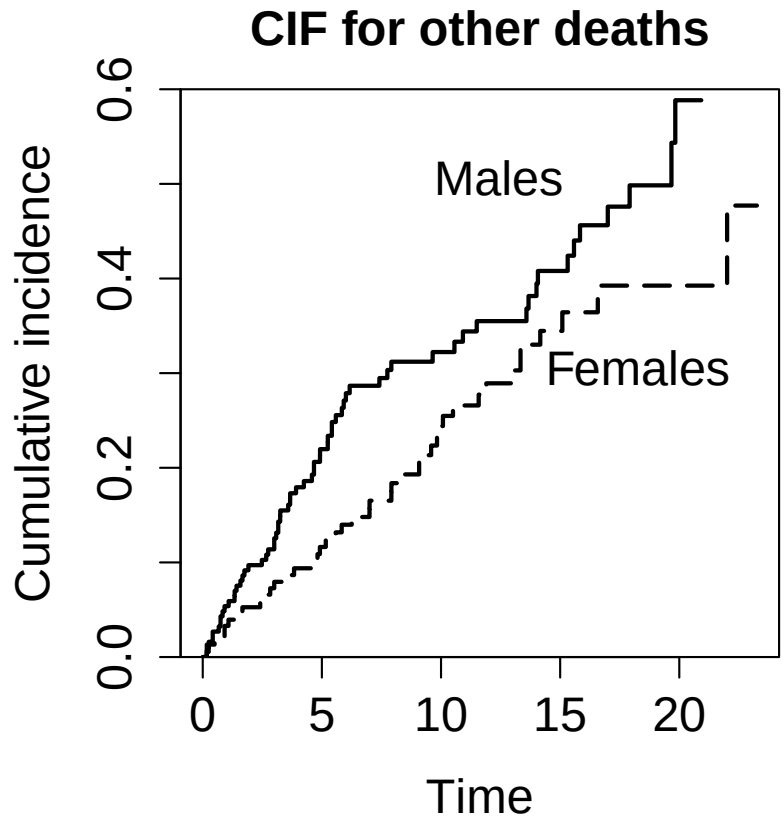
- AJ-estimates of CIFs (solid) for both causes.
- Naive KM-estimates of CIF (dashed) $>$ AJ-estimates
- CIF curves may also be stacked (right).



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).