Statistical Methods in Cancer Epidemiology using R

Karri Seppä

Finnish Cancer Registry

Lecture 8

karri.seppa@cancer.fi

Mar 23, 2020

Timetable today

- Lecture 8 (competing risks):
  - 9.15-10.00
- Practical 7 (survival without competing risks):
  - 10.15 ->
- Lecture 9 (population-based cancer survival):
  - 13.15-14.00
- Practical 8 (competing risks):
  - 14.15 ->

Points to be covered

1. Competing risks: event-specific cumulative incidences & hazards.

2. Kaplan–Meier and Aalen–Johansen estimators.

3. Regression modelling of competing hazards:

4. Packages `survival`, `mstate`, `cmprisk`.

5. Functions `Surv()`, `survfit()`, `plot.survfit()`, `coxph()`, `Cuminc()`.

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

```
orca <- read.table(file="oralca2.txt")
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 mortality

```
library(survival)
orca$suob <- Surv(orca$time, 1*(orca$event > 0) )
km1 <- survfit( suob ~ 1, data = orca)
km1                 #  brief  summary
```

```
Call: survfit(formula = suob ~ 1, data = orca)

      n  events  median 0.95LCL 0.95UCL
 338.00  229.00    5.42    4.33    6.92
```

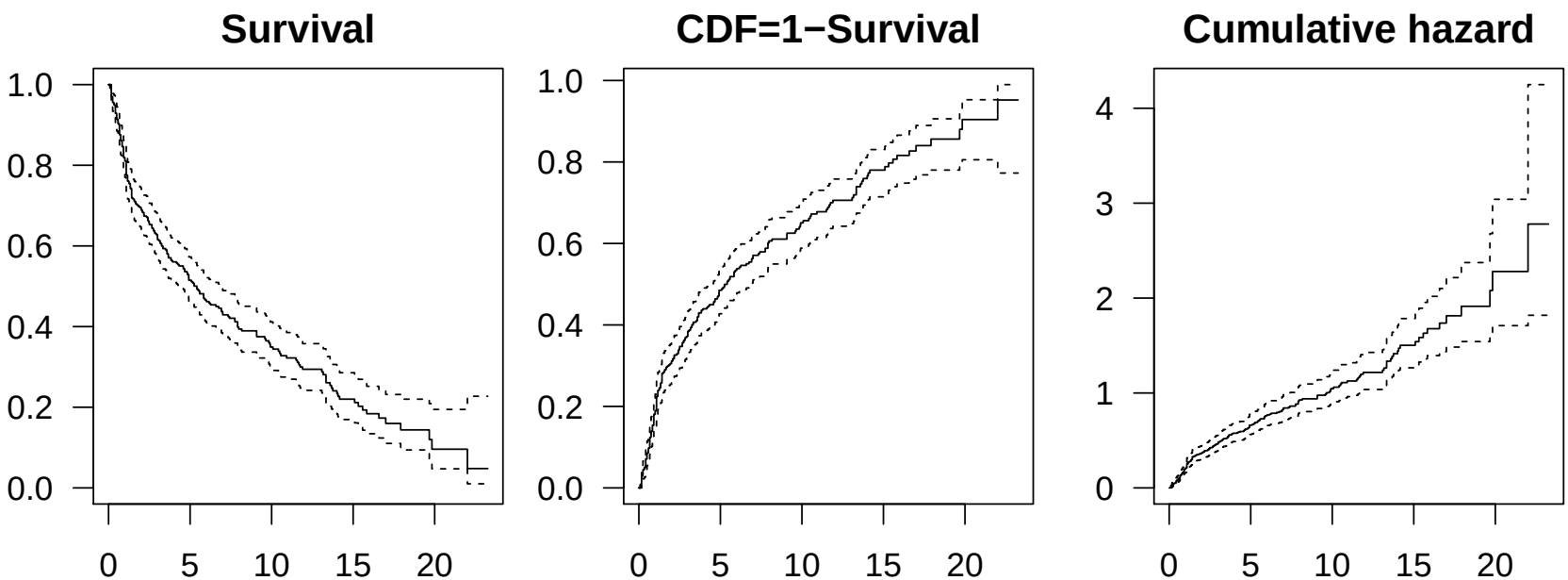
```
summary(km1)      #  detailed KM-estimate
```

```
Call: survfit(formula = suob ~ 1, data = orca)

   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
  0.249    327       1   0.9645 0.01007       0.9450        0.984
  0.252    326       3   0.9556 0.01120       0.9339        0.978
  0.329    323       1   0.9527 0.01155       0.9303        0.976
```

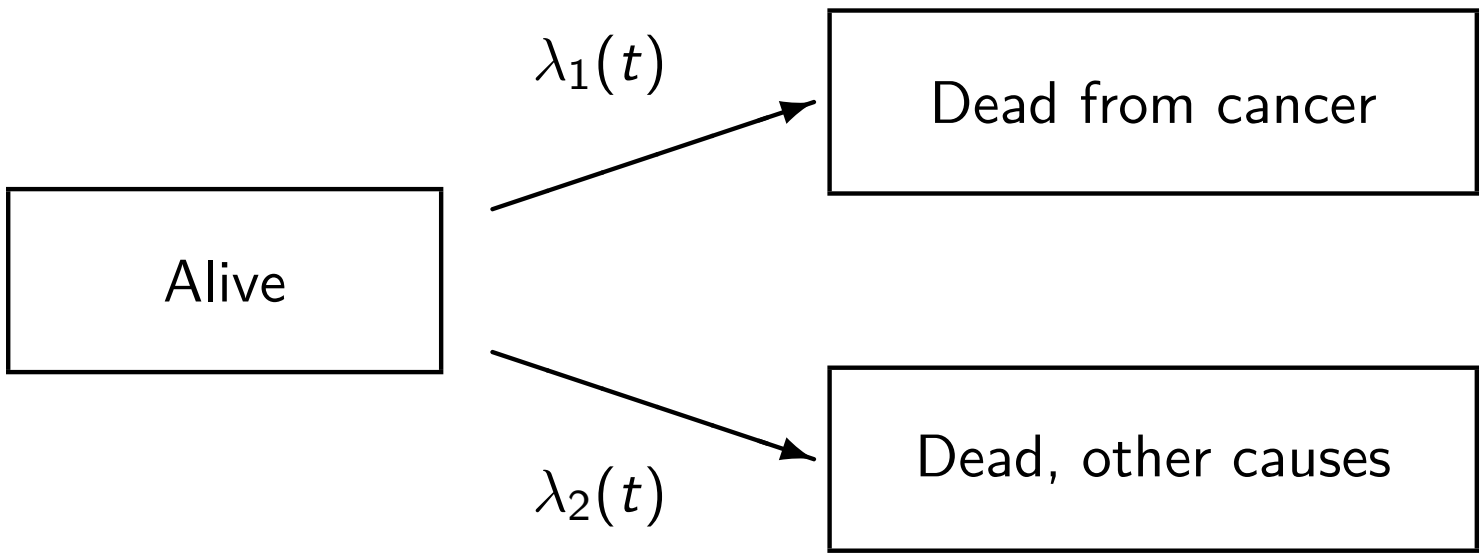
Oral cancer: Kaplan-Meier estimates

```
par(mfrow=c(1,3),mar=c(3,2.5,2,0.5),las=1)
plot(km1,main="Survival")
plot(km1,fun="F",main="CDF=1-Survival")
plot(km1,fun="cumhaz",main="Cumulative hazard")
```



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 **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 ($\Delta_k$) 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:

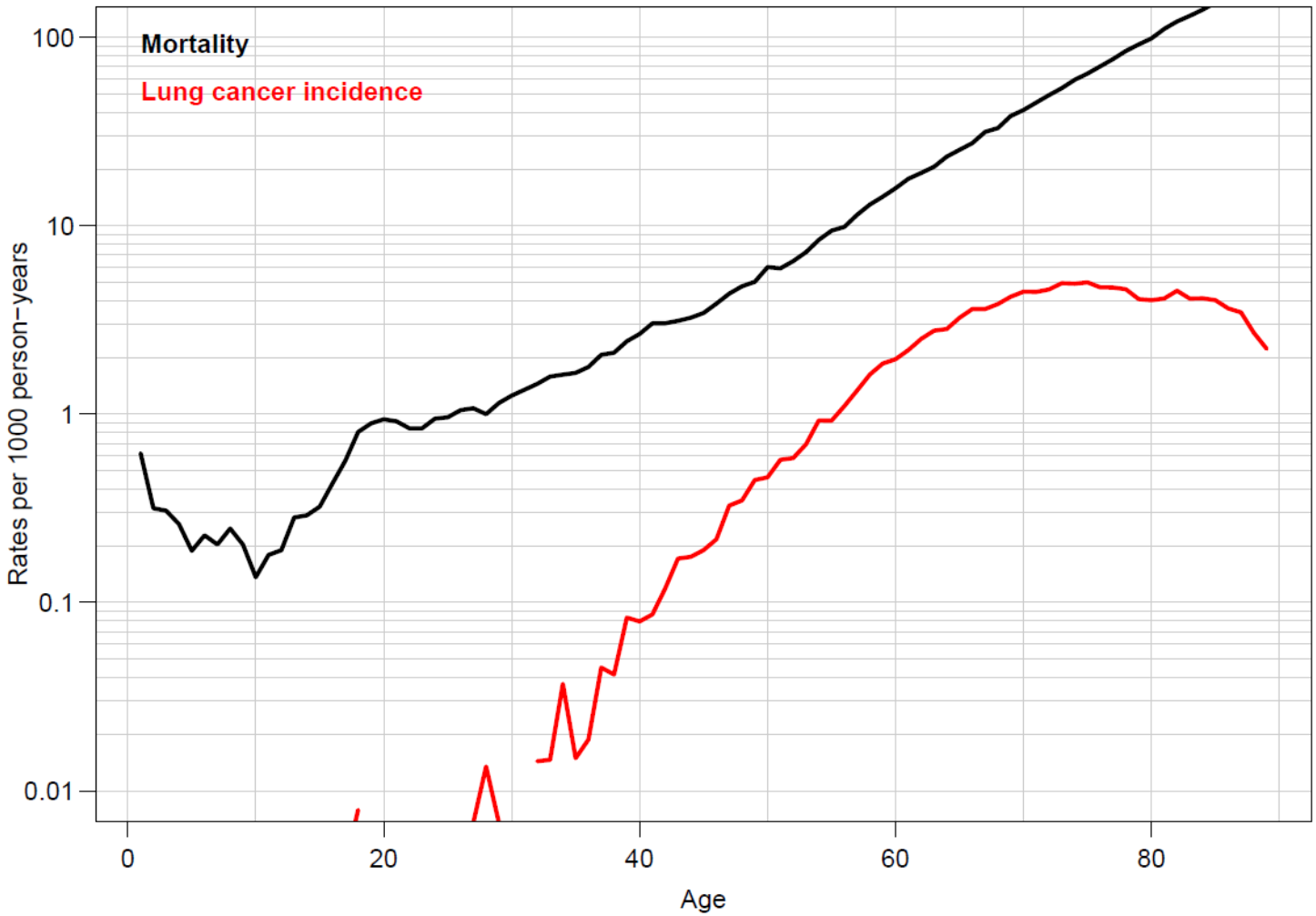
$$F_c^*(a) = 1 - \exp\{-\Lambda_c(a)\}.$$

- Yet, cancer occurs in a mortal population.

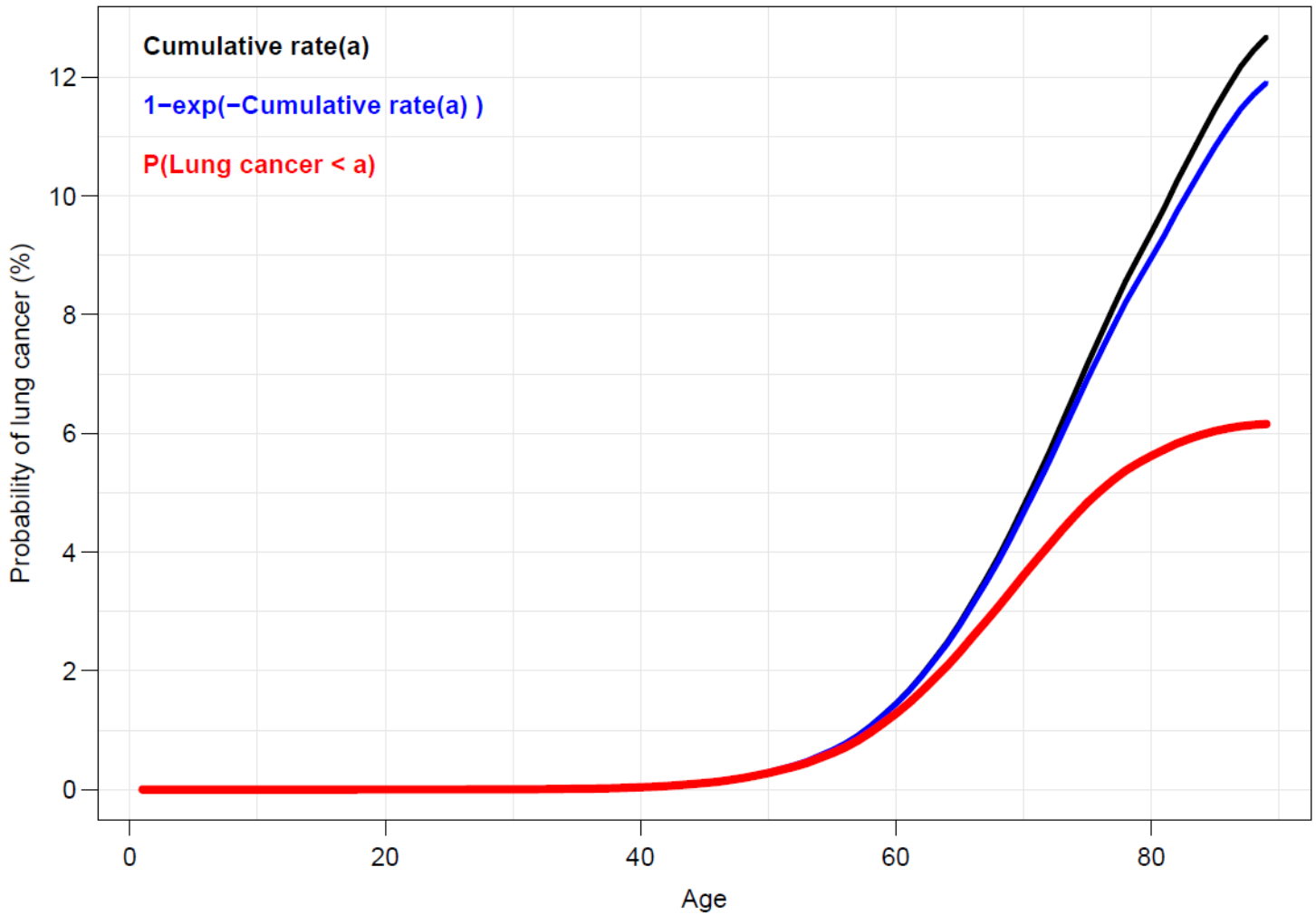
- As such CR(75) is a sound age-standardized summary measure for comparing cancer incidence across populations based on a neutral standard population.

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

Ex. Survival from oral cancer

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

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

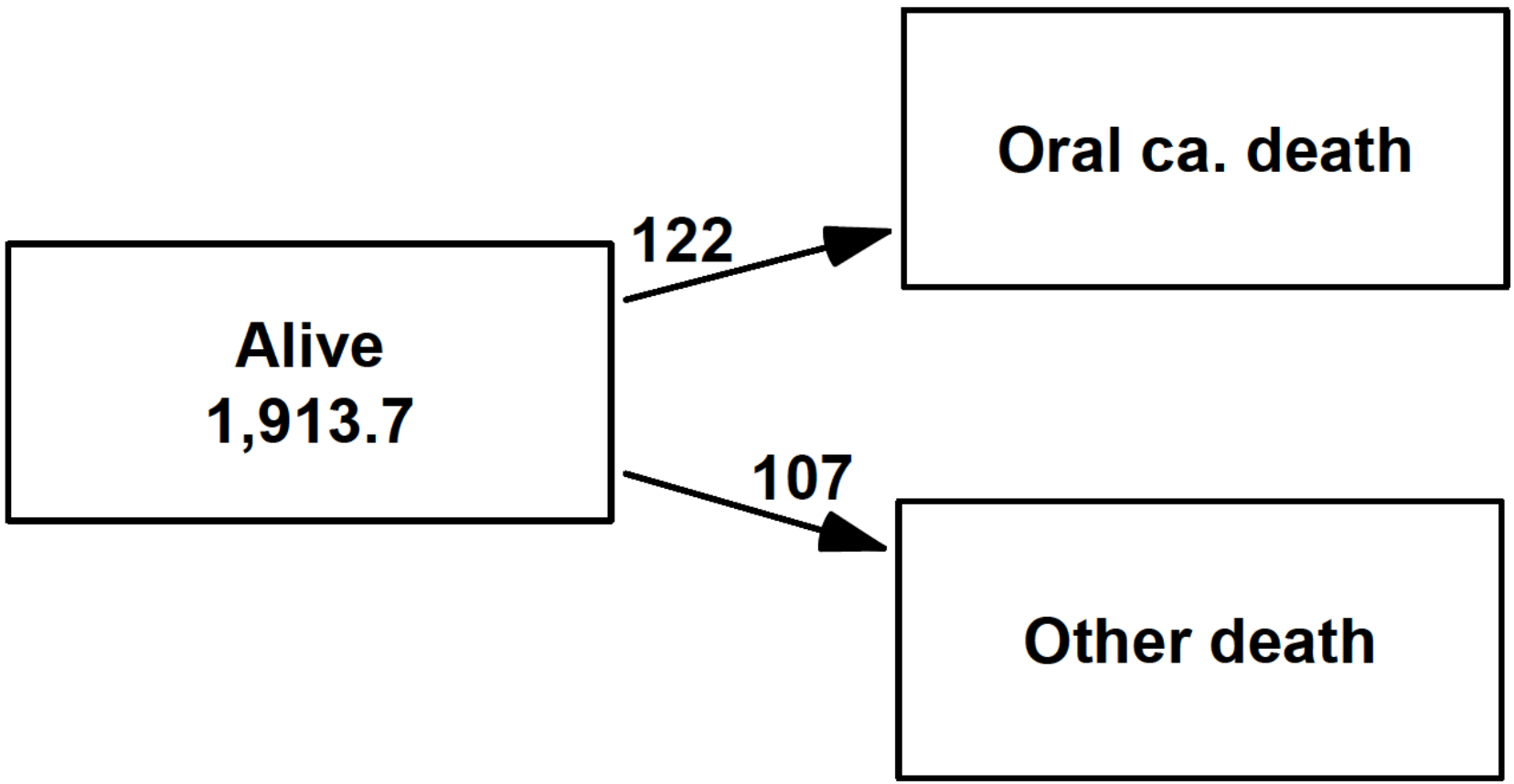
```
Transitions:
     To
From    Alive Oral ca. death Other death  Records:  Events: Risk time:
  Alive   109            122         107       338      229    1913.67

Transitions:
     To
From     Persons:
  Alive       338
```

Box diagram for transitions

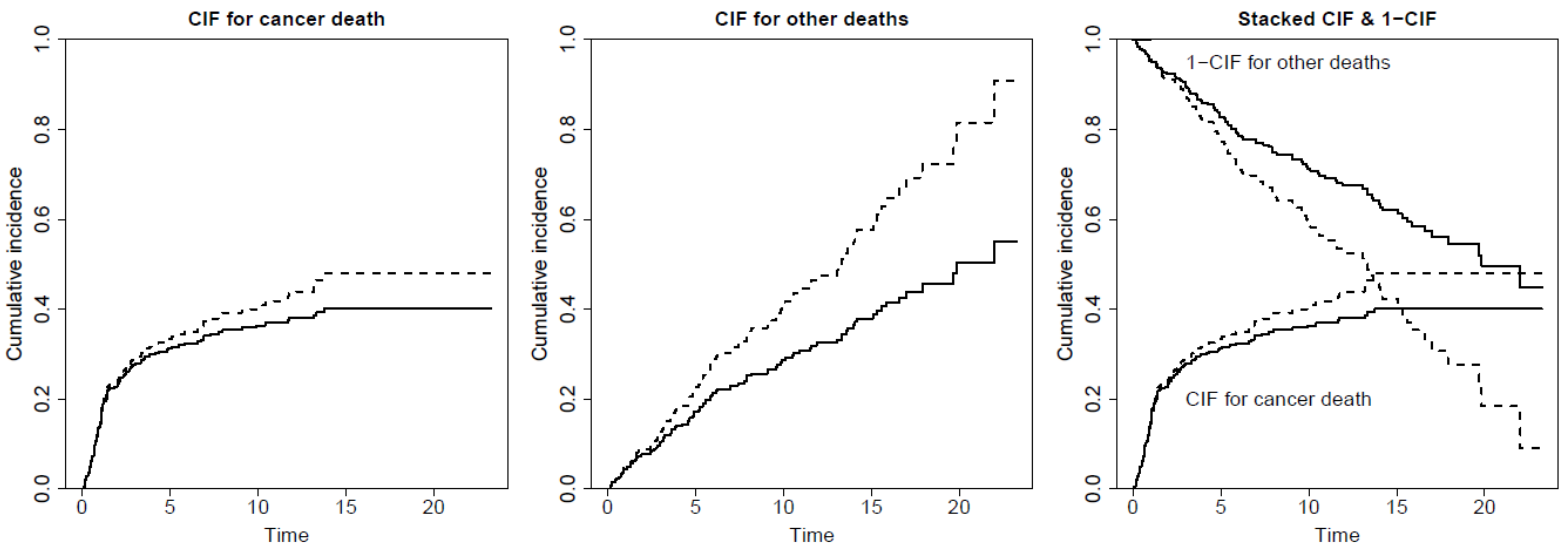
Interactive use of function `boxes()`.

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



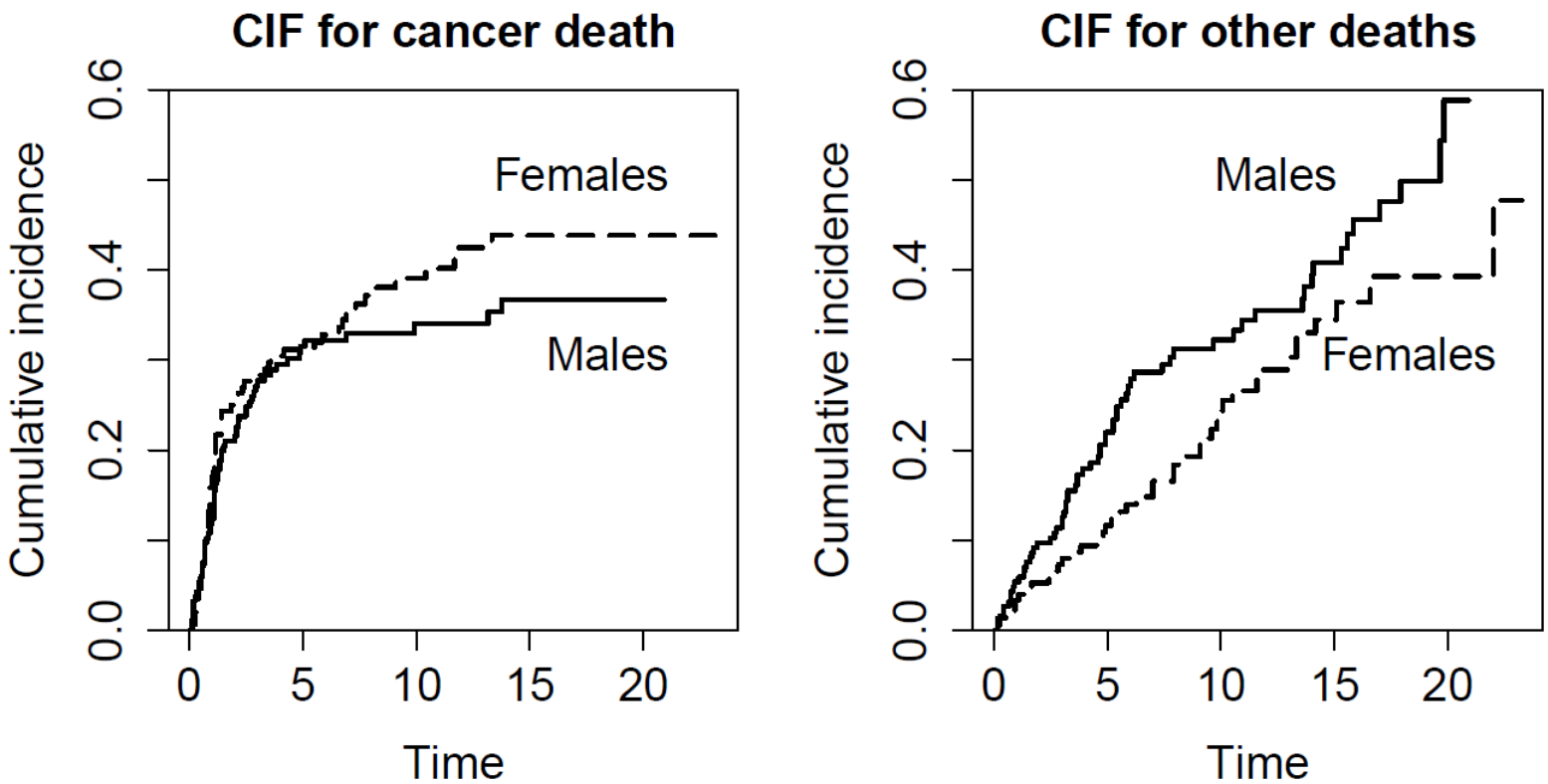
Ex. Survival from oral cancer

- 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).

Modelling with competing risks

Main options, providing answers to different questions:

(a) Cox model for event-specific hazards $\lambda_c(t) = f_c(t)/[1 - F(t)]$, when *e.g.* the interest is in the biological effect of the prognostic factors on the fatality of the very disease that often leads to the relevant outcome.

If $\lambda_1(y_i)$ and $\lambda_2(y_i)$ have no common parameters, they may be fitted separately treating competing events as censorings.
– Still, avoid estimating "net risks" from $F_c^* = 1 - \exp(-\Lambda_c)$!

(b) **Fine–Gray model** for the hazard of the subdistribution $\gamma_c(t) = f_c(t)/[1 - F_c(t)]$ when we want to assess the impact of the factors on the overall cumulative incidence of event c.
– Function `crr()` in package `cmprsk`.

Fine-Gray model

- Regression model for subdistribution hazards
- CIF quantifies cumulative disease incidence
- Risk ratio of subdistribution hazards can be used to assess the effects of covariates on CIF
- The same subdistribution hazard model allow one to make inferences about the relative magnitudes of the effects of the covariates on the incidence of the given type of event
- The direction of the subdistribution hazard ratio denotes the direction but does not directly provide the magnitude of the effect of the covariate on the CIF.

Estimation: Fine-Gray model vs Cox model

```
orca <- read.table("oralca2.txt", header=T)
orca$agegr <- cut(orca$age,c(0,55,75,Inf),right=F)
table(orca$agegr)
```

```
  [0,55)  [55,75) [75,Inf)
     100      160       78
```

- Estimation of cause-specific hazards

```
cox1 <- coxph(Surv(time, event==1) ~ sex + agegr + stage, data=orca)
cox2 <- coxph(Surv(time, event==2) ~ sex + agegr + stage, data=orca)
```

- Estimation of subdistribution hazards

```
library(cmprsk)
fg1 <- crr(orca$time, orca$event, cov1 = model.matrix(cox1), failcode=1)
fg2 <- crr(orca$time, orca$event, cov1 = model.matrix(cox2), failcode=2)
```

Estimation: RRs for death from cancer

```
round( ci.exp(cox1), 2)
```

```
              exp(Est.) 2.5% 97.5%
sexMale            0.99 0.69  1.44
agegr[55,75)       1.52 0.94  2.46
agegr[75,Inf)      3.23 1.92  5.44
stageII            1.66 0.77  3.55
stageIII           2.20 1.03  4.74
stageIV            4.34 2.09  9.02
stageunkn          2.96 1.38  6.33
```

```
round(summary(fg1, Exp=T)$conf.int[,-2], 2)
```

```
              exp(coef) 2.5% 97.5%
sexMale            0.90 0.63  1.30
agegr[55,75)       1.34 0.82  2.18
agegr[75,Inf)      2.49 1.48  4.18
stageII            1.78 0.85  3.69
stageIII           2.19 1.03  4.67
stageIV            4.13 2.02  8.44
stageunkn          2.71 1.28  5.77
```

Estimation: RRs for death from other causes

```
round( ci.exp(cox2), 2)
```

```
              exp(Est.) 2.5% 97.5%
sexMale            1.76 1.18  2.64
agegr[55,75)       2.39 1.44  3.99
agegr[75,Inf)      4.38 2.44  7.85
stageII            0.82 0.44  1.55
stageIII           1.05 0.55  1.98
stageIV            1.45 0.74  2.83
stageunkn          1.44 0.74  2.83
```

```
round(summary(fg2, Exp=T)$conf.int[,-2], 2)
```

```
              exp(coef) 2.5% 97.5%
sexMale            1.68 1.14  2.49
agegr[55,75)       1.91 1.17  3.11
agegr[75,Inf)      2.07 1.14  3.73
stageII            0.86 0.47  1.57
stageIII           0.96 0.53  1.74
stageIV            0.73 0.38  1.43
stageunkn          0.87 0.46  1.67
```

Recommendations for analysing competing risk data

- Cumulative incidence functions (CIFs) should be used to estimate the incidence of each of the different types of competing risks. **Do not use the Kaplan-Meier estimate** of the survival function for this purpose.
- Researchers need to decide whether the research objective:
  - Use the **Fine-Gray** subdistribution hazard model when the focus is on estimating **incidence** or predicting **prognosis** in the presence of competing risks.
  - Use the **cause-specific hazard** model (e.g. Cox model) when the focus is on addressing **etiologic** questions.
- In some settings, both types of regression models should be estimated for each of the competing risks to permit a full understanding of the effect of covariates on the incidence and the rate of occurrence of each outcome.