

Time-to-event analysis for registry data: an introduction

Brice Ozenne (brice.ozenne@nru.dk)

¹ Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

² Section of Biostatistics, Department of Public Health, University of Copenhagen.

February 28th, 2023 - Brain Drugs WP3

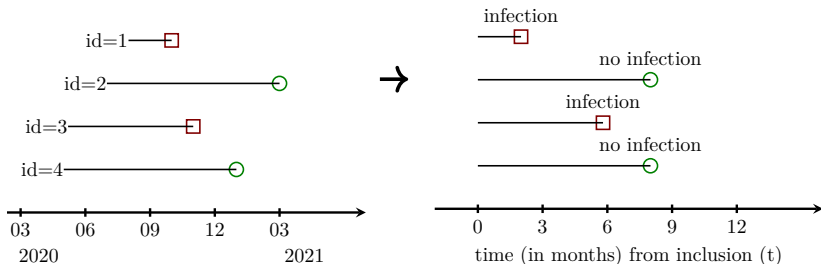
Defining a good target

- risk and rates as measures of disease frequency
 - risk/rates relationship
- time is important: from when? up to when?



Registry data as a cohort study

A group of n persons is followed over time



Two outcomes:

- $T_i \in [0, +\infty[$ time to event for subject i
(in months, or years, or ...)
- $\delta_i \in \{0, 1, 2\}$ type of event for subject i
(e.g. censoring, death due to COVID, death unrelated to COVID)

Typical study (1/2)

Find causes/remedies (E) to a disease/event:

- compare exposed and non-exposed
with respect to the frequency of the disease/event.
- interpretation and consequences

Description of event frequency:

Typical study (1/2)

Find causes/remedies (E) to a disease/event:

- compare exposed and non-exposed with respect to the frequency of the disease/event.
- interpretation and consequences

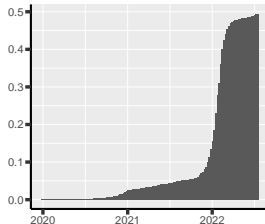
Description of event frequency:

- **risk**: proportion of people *getting* the event within a period τ

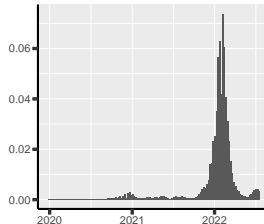
$$r(0; \tau) = \mathbb{P}[T \leq \tau, \delta = 1 | T > 0] \in [0, 1]$$

COVID example (1/2)

Risk of COVID infection
from 2019–12–30 in Denmark



1 week risk of COVID infection
in Denmark



Typical study

Find causes/remedies (E) to a disease/event:

- compare exposed and non-exposed with respect to the frequency of the disease/event.
- interpretation and consequences

Description of event frequency:

- **risk**: proportion of people *getting* the event within a period τ

$$r(0; \tau) = \mathbb{P}[T \leq \tau, \delta = 1 | T > 0] \in [0, 1]$$

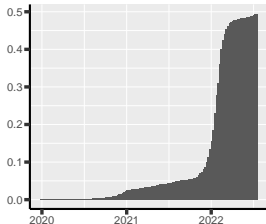
- **incidence rate**: risk of the event divided by at risk time

$$\lambda(t; \tau) = \frac{\mathbb{P}[T \leq t + \tau, \delta = 1 | T > t]}{\tau} \in [0, +\infty[$$

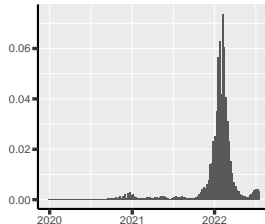
⚠ unit: time^{-1}

COVID example (2/2)

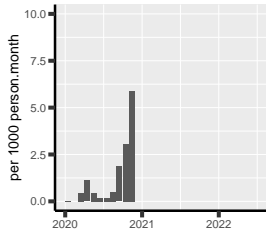
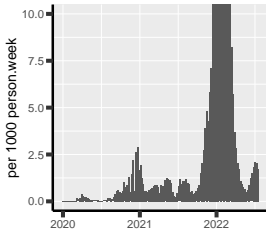
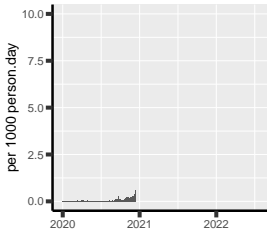
Risk of COVID infection
from 2019–12–30 in Denmark



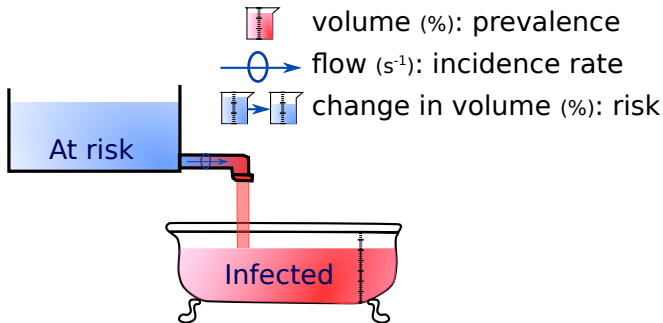
1 week risk of COVID infection
in Denmark



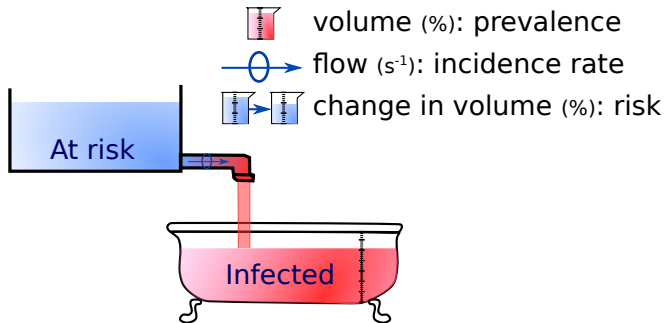
Incidence rate of COVID infection in Denmark



Risk-rate relationship



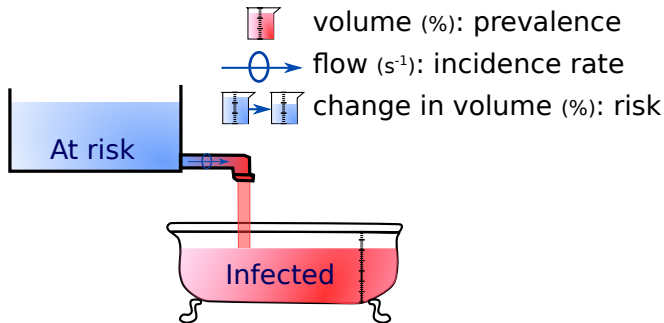
Risk-rate relationship



- *instantaneous* rate is also call hazard

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}[T \leq t + dt, N(t + dt) = 1 | T > t]}{dt}$$

Risk-rate relationship



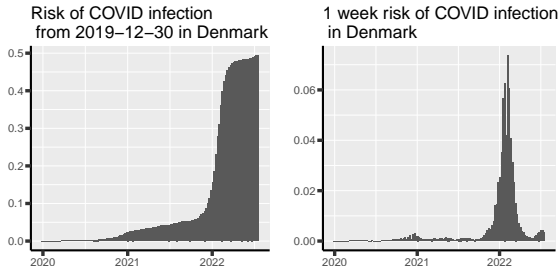
- *instantaneous* rate is also call hazard

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}[T \leq t + dt, N(t + dt) = 1 | T > t]}{dt}$$

- the risk can be deduced from the cumulating the hazard over the appropriate time interval

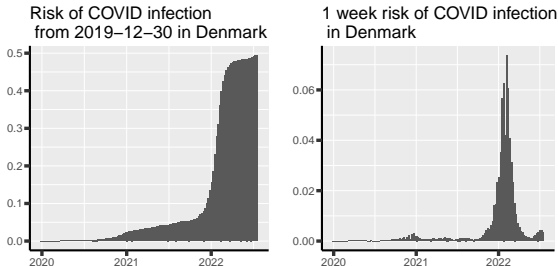
Definition of the parameter of interest

In many medical applications we are interest in the risk



Definition of the parameter of interest

In many medical applications we are interest in the risk

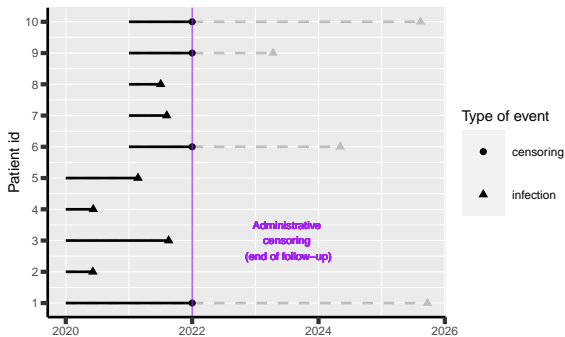


! there is no such thing as "the risk"

- of what? (e.g. COVID infection, death, ...)
- from when? (e.g. 01-01-2020, age 18, cancer diagnosis, ...)
- over which time period? (e.g. 1 week, 1 year, ...)

Example

Risk of death between start and end of follow-up: 53.4%

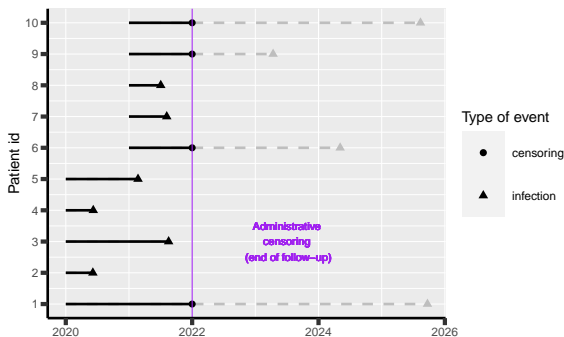


Example

Risk of death between start and end of follow-up: 53.4%



no clear interpretation! Mix of 5 year risk (42.5%)
and 10 year risk (64.2%)

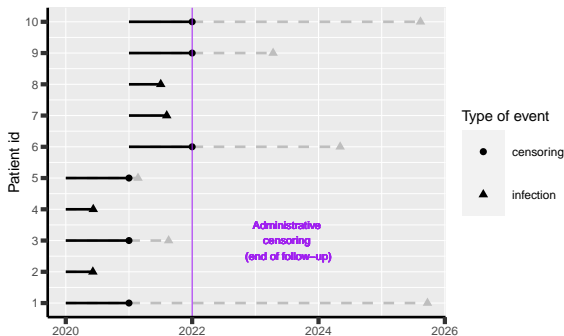


Example

Risk of death between start and end of follow-up: 53.4%



no clear interpretation! Mix of 5 year risk (42.5%)
and 10 year risk (64.2%)



Instead we could look at a specific time horizon (e.g. 1 year)

- censor events after this time

Time origin (Andersen et al., 2021)

"The follow-up time T_i is measured:

- from a meaningful starting point of the process (time 0)

which should be:

- unambiguously defined and comparable between individuals
- ideally clinically relevant."

"The choice of time origin should depend on the scientific questions" (and not the other way around)

Time origin (Andersen et al., 2021)

"The follow-up time T_i is measured:

- from a meaningful starting point of the process (time 0)

which should be:

- unambiguously defined and comparable between individuals
- ideally clinically relevant."

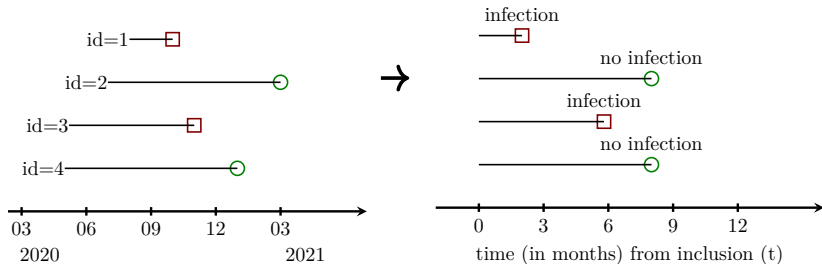
"The choice of time origin should depend on the scientific questions" (and not the other way around)



There may be several time scale:

- age
- time since diagnosis
- time since treatment initiation.

Time origin - in practice

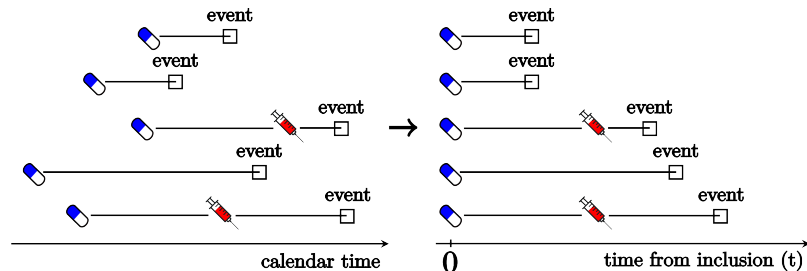


is "time from inclusion" meaningful?

- yes (time since diagnosis, time since treatment initiation)
- no (time since first participation to a research project)
→ age may be a better time scale

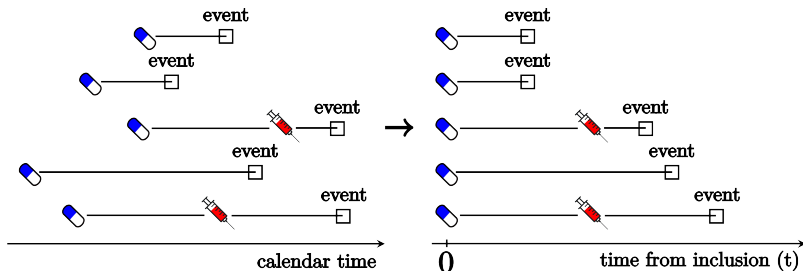
Exposure

With registry data, the exposure (often) vary over time



Exposure

With registry data, the exposure (often) vary over time



We can ask many different research questions:

- drug A vs. drug B (from baseline)
- drug A vs. A then B after 6 months
- drug A vs. A then B if A seems not effective
- ...

Analysis in an ideal world

- risk and rates calculations
 - G-formula
 - challenges



no censoring

no delayed entry

no confounders

no competing risks

fixed exposure

Estimation in an ideal word

- **risk:** proportion of people *getting* the event within a period τ

$$r(0; \tau) = \mathbb{P}[T \leq \tau, \delta = 1 | T > 0] \in [0, 1]$$

$$\hat{r}(0; \tau) = \frac{\text{"number of new cases"}}{\text{"number of persons at risk"}}$$

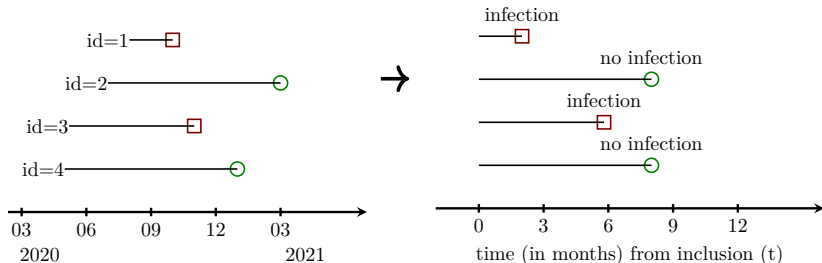
- **incidence rate:** risk of the event divided by at risk time

$$\lambda(0; \tau) = \frac{\mathbb{P}[T \leq \tau, \delta = 1 | T > 0]}{\tau} \in [0, +\infty[$$

$$\hat{\lambda}(0; \tau) = \frac{\text{"number of new cases"}}{\text{"cumulative at-risk time"}}$$

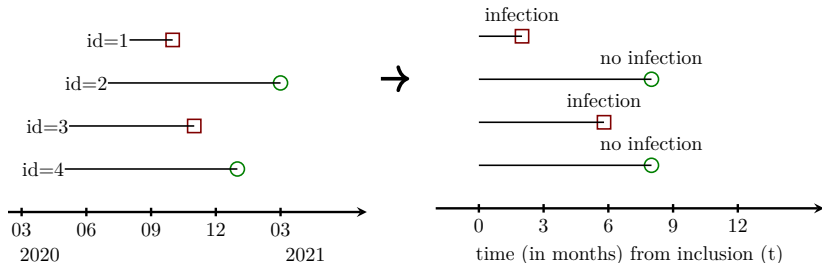
⚠ unit: time^{-1}

Toy example (risk)



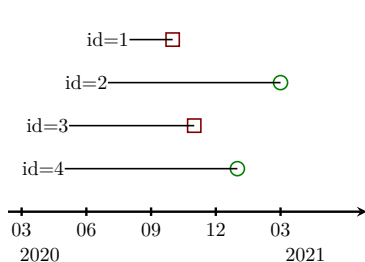
- $\hat{r}(0)$ = at baseline
- $\hat{r}(3)$ = after 3 months
- $\hat{r}(8)$ = after 8 months

Toy example (risk)



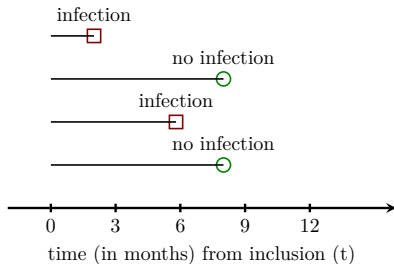
- $\hat{r}(0) = 0$ at baseline
- $\hat{r}(3) = 1/4$ after 3 months
- $\hat{r}(8) = 2/4$ after 8 months

Toy example (rate)



- $\tilde{T}_1 = 2$ months, $\tilde{Y}_1 = 1$
- $\tilde{T}_2 = 8$ months, $\tilde{Y}_2 = 0$

$\hat{\lambda}_\tau =$



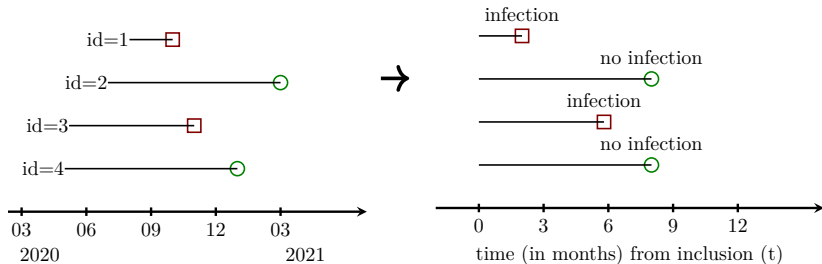
- $\tilde{T}_3 = 5.9$ months, $\tilde{Y}_3 = 1$
- $\tilde{T}_4 = 8$ months, $\tilde{Y}_4 = 0$

\approx per person-month

\approx per 1000 person-month

\approx per person-year

Toy example (rate)



- $\tilde{T}_1 = 2$ months, $\tilde{Y}_1 = 1$

- $\tilde{T}_2 = 8$ months, $\tilde{Y}_2 = 0$

- $\tilde{T}_3 = 5.9$ months, $\tilde{Y}_3 = 1$

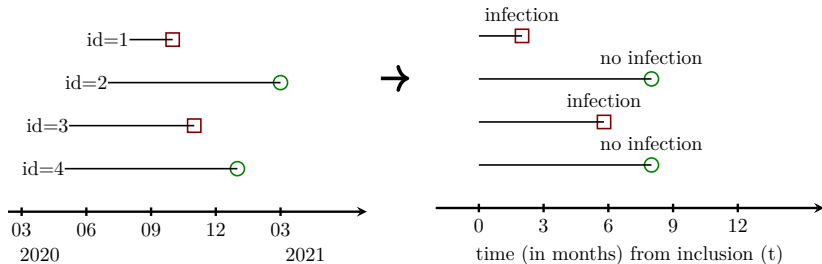
- $\tilde{T}_4 = 8$ months, $\tilde{Y}_4 = 0$

$$\hat{\lambda}_\tau = \frac{1 + 0 + 1 + 0}{2 + 8 + 5.9 + 8} = \frac{2 \text{ new cases}}{23.8 \text{ person-month}} \approx 0.084 \text{ per person-month}$$

$$\approx 84 \text{ per 1000 person-month}$$

$$\approx \text{per person-year}$$

Toy example (rate)



- $\tilde{T}_1 = 2$ months, $\tilde{Y}_1 = 1$

- $\tilde{T}_3 = 5.9$ months, $\tilde{Y}_3 = 1$

- $\tilde{T}_2 = 8$ months, $\tilde{Y}_2 = 0$

- $\tilde{T}_4 = 8$ months, $\tilde{Y}_4 = 0$

$$\hat{\lambda}_\tau = \frac{1 + 0 + 1 + 0}{2 + 8 + 5.9 + 8} = \frac{2 \text{ new cases}}{23.8 \text{ person-month}} \approx 0.084 \text{ per person-month}$$

$$\approx 84 \text{ per 1000 person-month}$$

$$\frac{2 \text{ new cases}}{23.8/12 \text{ person-year}} \approx 1.004 \text{ per person-year}$$

What about heterogeneity in treatment effect?

Vaccination of children of different ages:

| | age | [-1,10] | (10,120] | (120,300] |
|--------------|-----|--------------|---------------|---------------|
| bcg status | | | | |
| no censored | | 238 (94.07%) | 1268 (95.05%) | 370 (95.85%) |
| dead | | 15 (5.93%) | 66 (4.95%) | 16 (4.15%) |
| yes censored | | 30 (100%) | 1790 (96.91%) | 1356 (95.22%) |
| dead | | 0 (0%) | 57 (3.09%) | 68 (4.78%) |
| risk | | | | |
| difference | | -5.929 | -1.861 | 0.63 |
| ratio | | 0 | 0.624 | 1.152 |

What about heterogeneity in treatment effect?

Vaccination of children of different ages:

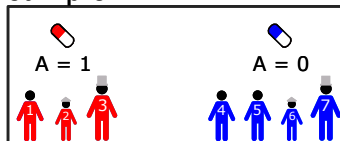
| | age | [-1,10] | (10,120] | (120,300] |
|------------|----------|--------------|---------------|---------------|
| bcg status | | | | |
| no | censored | 238 (94.07%) | 1268 (95.05%) | 370 (95.85%) |
| | dead | 15 (5.93%) | 66 (4.95%) | 16 (4.15%) |
| yes | censored | 30 (100%) | 1790 (96.91%) | 1356 (95.22%) |
| | dead | 0 (0%) | 57 (3.09%) | 68 (4.78%) |
| risk | | | | |
| difference | | -5.929 | -1.861 | 0.63 |
| ratio | | 0 | 0.624 | 1.152 |

- model and report the age-specific effect $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$
- ~~model a constant effect and report this effect~~
- model the age-specific effect and report a standardized effect

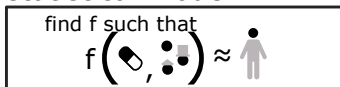
$$\hat{\Psi} = f(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$$

Intuition behind standardization

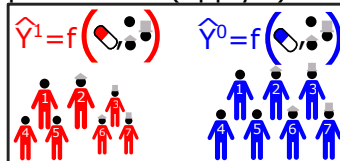
sample



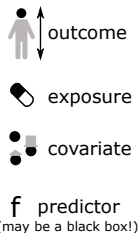
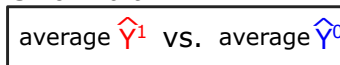
statistical model



predictions (apply f)




G-formula



Standardization in practice (aka G-formula)

2 equivalent implementations:

- predictions, e.g. `riskRegression::ate` function in 
- weighted average of the strata-specific effects

$$\psi = \theta_1 \mathbb{P}(\text{age} \in (0, 10]) + \theta_2 \mathbb{P}(\text{age} \in (10, 120]) + \theta_3 \mathbb{P}(\text{age} \in (120, 212])$$

Here for the risk difference:

$$\psi = -5.929 \frac{269}{5274} - 1.861 \frac{3181}{5274} + 0.630 \frac{1810}{5274} = -1.22$$

Exercise!

Open the file `exercise-workshopEpi.R` (line 18-97)

Load data the bissau dataset:

- visualize the individual survival trajectories
- compare the risk per vaccine group accounting or not for age

⚠ to avoid data management we will do what we should not do:

- ignore difference in at risk time/right censoring,
i.e. assume that children who left early the study will not die
by 183 days (max follow-up time)
→ systematic underestimation of the risk!

⚠ age groups are artificial

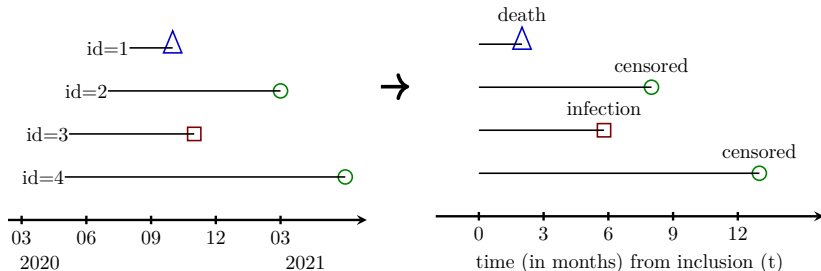
Challenge 1: partially observed outcome

(a) competing risks (death or other brain disorders):

- prevent occurrence of the event of interest

(b) right-censoring:

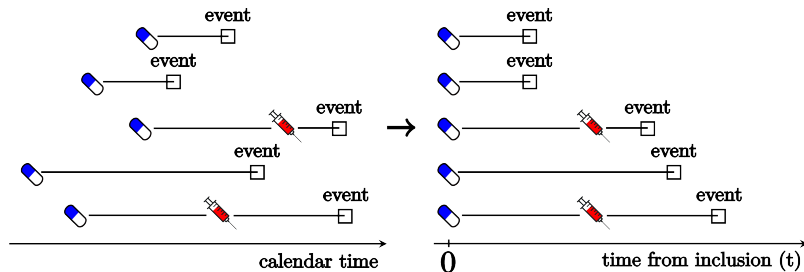
- event may or may not have occurred after last observation



Can we exclude dead/censored patients?

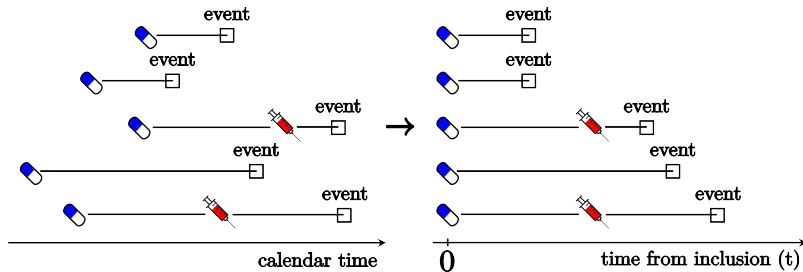
Consider dead patients as free of infection?

Challenge 2: time-varying exposure



Can we compare never switchers to switchers?

Challenge 2: time-varying exposure



Can we compare never switchers to switchers?

→ ECF presentation (20/10/2022)

Principles (Andersen and Keiding, 2012)

(1) Do not condition on the future

- ✗ Use future information to exclude patients
- ✗ Use future information to decide on past exposure

(2) Do not condition on having reached an absorbing state

- ✗ Consider dead patients to be at risk of stroke (death as no event)
- ✗ Model biomarker values of dead patients

(3) Stick to this world

- ✗ Consider a world where patients do not die (death as censoring)
"if you do not die within a year, this treatment is beneficial ..."

Many other challenges (Pazzagli et al., 2018)

Definition of the exposure:

- reconstruction of the exposure based on purchasing dates

Time-varying confounding

- confounder variables may change over time
... due to the exposure → cannot use 'traditional adjustment'
(e.g. CD4 counts when studying HIV treatments)

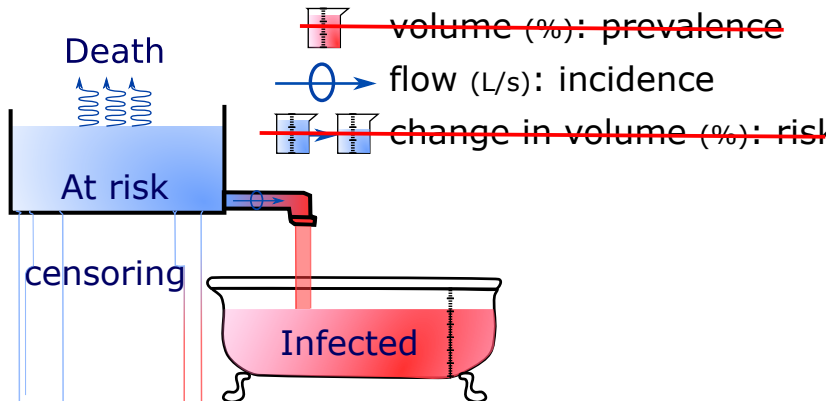
Complex exposure:

- patient may switch exposure, often for health-related reason
- the exposure is not binary but may be time or dose related

Big picture

Because of complications we will (often) model the incidence

- and then deduce the risk

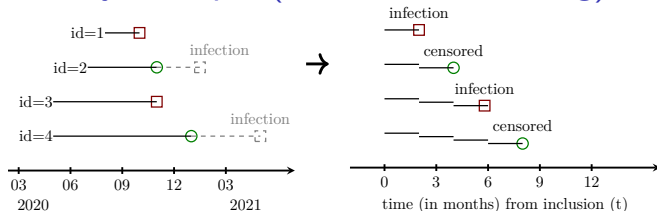


⚠ do not loose track of what you want because of a detour!

Handling censoring

- Kaplan Meier (KM)
- independent censoring assumption
 - revisiting KM as a weighting approach

Toy example (risk under censoring)



Risk after 8 months:

- $\hat{r}(8) =$

Incidence:

- $\hat{\lambda}_1 =$
- $\hat{\lambda}_2 =$
- $\hat{\lambda}_3 =$
- $\hat{\lambda}_4 =$

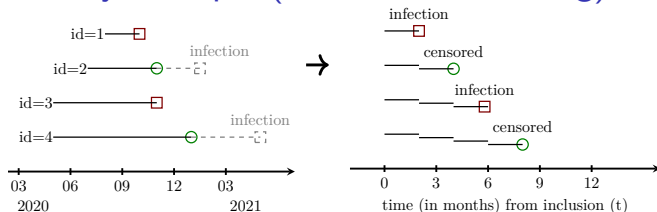
$$t \in [0; 2]$$

$$t \in [2; 4]$$

$$t \in [4; 6]$$

$$t \in [6; 8]$$

Toy example (risk under censoring)



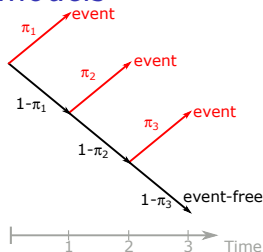
Risk after 8 months:

- $\hat{r}(8) = (2+?)/4 = 0.5 \text{ or } 0.75$

Incidence:

- $\hat{\lambda}_1 = 1/(2 + 2 + 2 + 2) = 1/8$ $t \in [0; 2]$
- $\hat{\lambda}_2 = 0/(2 + 2 + 2) = 0$ $t \in [2; 4]$
- $\hat{\lambda}_3 = 1/(2 + 2) = 1/4$ $t \in [4; 6]$
- $\hat{\lambda}_4 = 0/2 = 0$ $t \in [6; 8]$

Binary probability models



Survival (probability of not getting the event)

$$S(3) = \mathbb{P}[T > 3] = \mathbb{P}[T > 1] \mathbb{P}[T > 2 | T > 1] \mathbb{P}[T > 3 | T > 2]$$

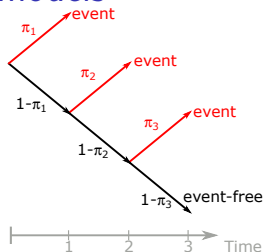
$$=$$

Risk (probability of getting the event)

$$r(3) = \mathbb{P}[T \leq 3] =$$

$$=$$

Binary probability models



Survival (probability of not getting the event)

$$\begin{aligned} S(3) &= \mathbb{P}[T > 3] = \mathbb{P}[T > 1] \mathbb{P}[T > 2 | T > 1] \mathbb{P}[T > 3 | T > 2] \\ &= (1 - \pi_1)(1 - \pi_2)(1 - \pi_3) \end{aligned}$$

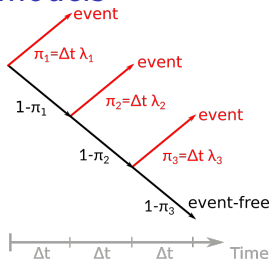
Risk (probability of getting the event)

$$\begin{aligned} r(3) &= \mathbb{P}[T \leq 3] = 1 - S(3) = 1 - (1 - \pi_1)(1 - \pi_2)(1 - \pi_3) \\ &= \end{aligned}$$

Binary probability models

Assuming piecewise constant hazard:

- $\pi_t = \Delta t \lambda_t$: disease frequency equals rate times duration in each time interval



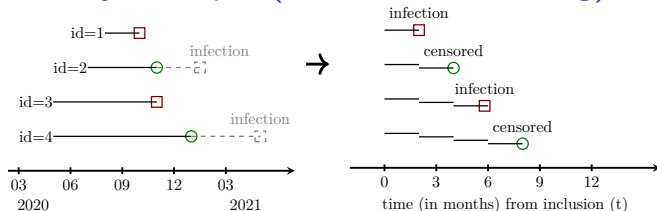
Survival (probability of not getting the event)

$$\begin{aligned} S(3) &= \mathbb{P}[T > 3] = \mathbb{P}[T > 1] \mathbb{P}[T > 2 | T > 1] \mathbb{P}[T > 3 | T > 2] \\ &= (1 - \pi_1)(1 - \pi_2)(1 - \pi_3) \end{aligned}$$

Risk (probability of getting the event)

$$\begin{aligned} r(3) &= \mathbb{P}[T \leq 3] = 1 - S(3) = 1 - (1 - \pi_1)(1 - \pi_2)(1 - \pi_3) \\ &= 1 - (1 - \Delta t \lambda_1)(1 - \Delta t \lambda_2)(1 - \Delta t \lambda_3) \end{aligned}$$

Toy example (risk under censoring)



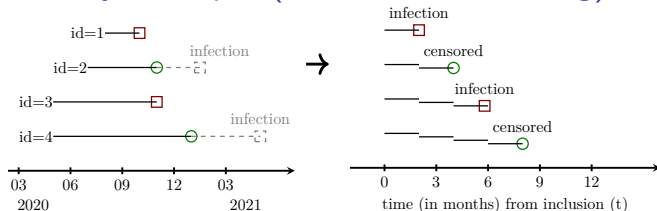
Risk after 8 months:

- $\hat{r}(8) =$
- $\hat{r}(8) = 1 - (1 - \hat{\lambda}_1 \Delta t_1)(1 - \hat{\lambda}_2 \Delta t_2)(1 - \hat{\lambda}_3 \Delta t_3)(1 - \hat{\lambda}_4 \Delta t_4)$
 $= 1 - (1 - 1/8 * 2) * 1 * (1 - 1/4 * 2) * 1 = 0.625$

Incidence:

- $\hat{\lambda}_1 = 1/8$ $t \in [0; 2]$
- $\hat{\lambda}_2 = 0$ $t \in [2; 4]$
- $\hat{\lambda}_3 = 1/4$ $t \in [4; 6]$
- $\hat{\lambda}_4 = 0$ $t \in [6; 8]$

Toy example (risk under censoring)



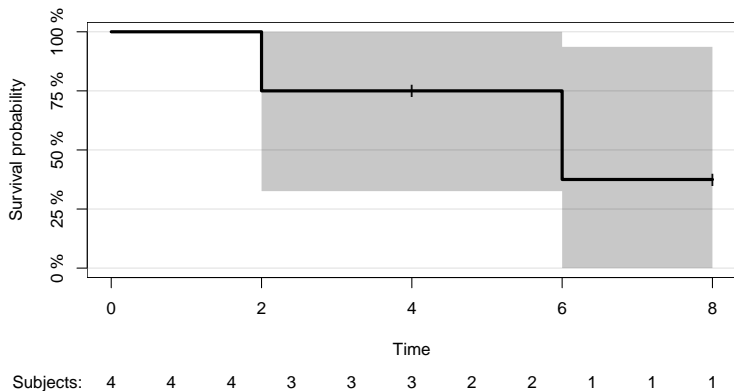
Risk after 8 months:

- $\hat{r}(8) = (2+?)/4 = 0.5$ or 0.75
- $\hat{r}(8) = 1 - (1 - \hat{\lambda}_1 \Delta t_1)(1 - \hat{\lambda}_2 \Delta t_2)(1 - \hat{\lambda}_3 \Delta t_3)(1 - \hat{\lambda}_4 \Delta t_4)$
 $= 1 - (1 - 1/8 * 2) * 1 * (1 - 1/4 * 2) * 1 = 0.625$

Incidence:

- $\hat{\lambda}_1 = 1/8$ $t \in [0; 2]$
- $\hat{\lambda}_2 = 0$ $t \in [2; 4]$
- $\hat{\lambda}_3 = 1/4$ $t \in [4; 6]$
- $\hat{\lambda}_4 = 0$ $t \in [6; 8]$

Kaplan Meier in R



```
library(prodlm)
e.KM <- prodlm(Hist(time,event) ~ 1, data = df)
plot(e.KM, marktime = TRUE)
```


Independent censoring assumption

For a patient alive at time t , his censoring status should not be informative of his risk of infection at any later timepoint.

- patients who stay are similar (i.e. representative) of those who are lost to follow-up
- ✓ administrative censoring (end of study)
- ✗ health-related censoring
(subject was so sick so he had to leave the study)
(subject is not fearing to catch the disease anymore)

Exercise!

Open the file `exercise-workshopEpi.R` (line 98-156)

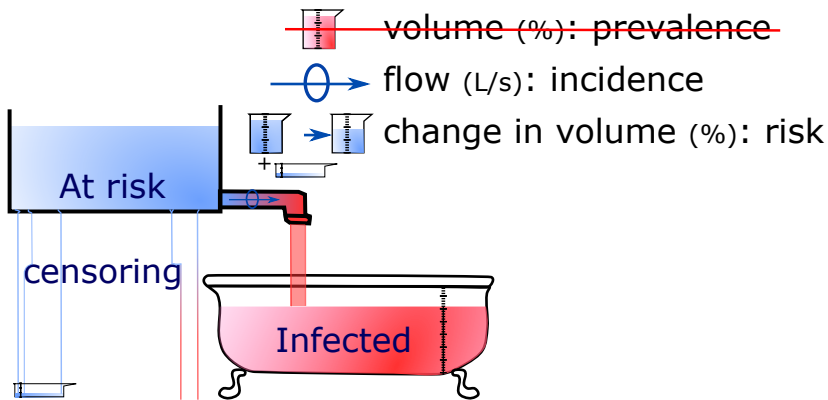
Run the code used to analyse the toy example

Re-analyze the data from the Bissau study, accounting for right-censoring

- how is the estimated risk compared to 'ignoring censoring'?
- what key assumption are we making?

Another point of view

Recover the risk based on the censoring process (instead of the rate)



IPCW point of view (Satten and Datta, 2001)

Without censoring we could estimate the survival at time t by:

$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_i \leq t}$$

where T_i is the time to event for individual i .

IPCW point of view (Satten and Datta, 2001)

Without censoring we could estimate the survival at time t by:

$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_i \leq t}$$

where T_i is the time to event for individual i .

We now also consider C_i , the time to censoring.

$\delta_i \in \{0, 1\}$ indicates whether censoring or event is observed.

- censored observations at time t will not contribute
- uncensored observations at time t will contribute, weighted by the inverse of their probability to be observed.

$$\hat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_{T_i \leq t} \delta_i}{\mathbb{P}[C_i \geq t]}$$

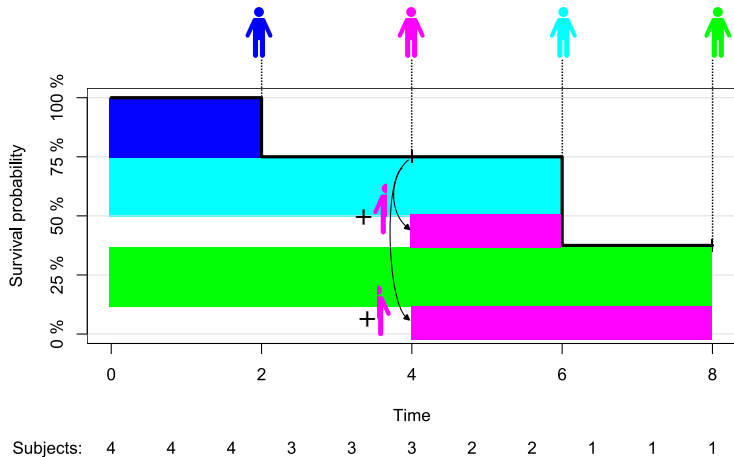
Target
○○○○○○○
○○○○○

Ideal world
○○○○○○○
○○○○○

Handling censoring
○○○○○○○
○○●

Reference
○○○

Efron redistribution algorithm



Reference I

- Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12):1074–1088.
- Andersen, P. K., Pohar Perme, M., van Houwelingen, H. C., Cook, R. J., Joly, P., Martinussen, T., Taylor, J. M. G., Abrahamowicz, M., and Therneau, T. M. (2021). Analysis of time-to-event for observational studies: Guidance to the use of intensity models. *Statistics in Medicine*, 40(1):185–211.
- Jensen, H., Benn, C. S., Lisse, I. M., Rodrigues, A., Andersen, P. K., and Aaby, P. (2007). Survival bias in observational studies of the impact of routine immunizations on childhood survival. *Tropical Medicine & International Health*, 12(1):5–14.

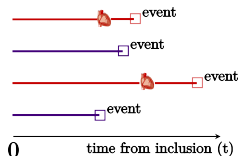
Reference II

Pazzagli, L., Linder, M., Zhang, M., Vago, E., Stang, P., Myers, D., Andersen, M., and Bahmanyar, S. (2018). Methods for time-varying exposure related problems in pharmacoepidemiology: an overview. *Pharmacoepidemiology and drug safety*, 27(2):148–160.

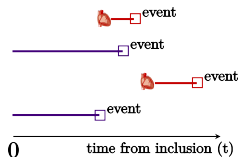
Satten, G. A. and Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.

Immortal time bias (1/2)

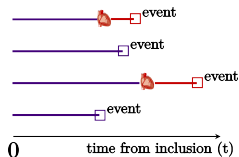
Solution 1:



Solution 2:



Solution 3:

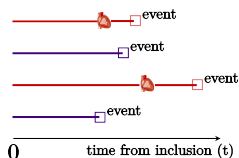


Immortal time bias (1/2)

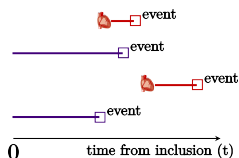


Solution 1:

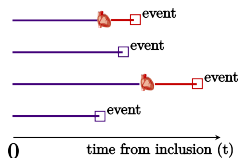
- unrealistic: use future information to define exposure
- immortal time bias: baseline-transplant



Solution 2:



Solution 3:

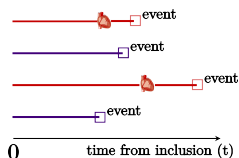


Immortal time bias (1/2)



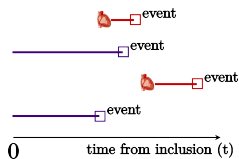
Solution 1:

- unrealistic: use future information to define exposure
- immortal time bias: baseline-transplant

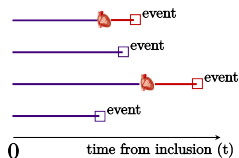


Solution 2:

- unrealistic: use future information to remove data
- biased against no transplant



Solution 3:

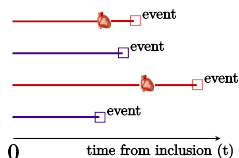


Immortal time bias (1/2)



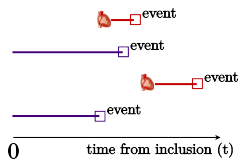
Solution 1:

- unrealistic: use future information to define exposure
- immortal time bias: baseline-transplant



Solution 2:

- unrealistic: use future information to remove data
- biased against no transplant

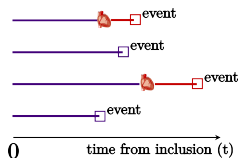


Solution 3:

- realistic: time-varying exposure



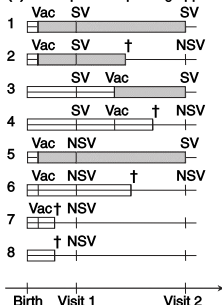
how to carry-out the analysis?



Immortal time bias (2/2)

From Jensen et al. (2007):

(a) Retrospective updating approach



SV = Seen vaccination card

NSV = Not seen vaccination card

□ = classified as unvaccinated

■ = classified as vaccinated

Vac = vaccinated, † = dead.

Retrospective updating approach

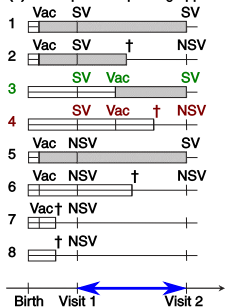
In the retrospective updating approach, vaccination status is used as a time-varying variable changing from unvaccinated to vaccinated, on the *exact date of vaccination*. This is a standard statistical approach if vaccination information is collected for all children, regardless of survival status.

While for the retrospective updating approach, the vaccination status is updated on the exact date of vaccination, regardless of survival status, the prospective updating approach updates the vaccination status on the date of the last visit for the child.

Immortal time bias (2/2)

From Jensen et al. (2007):

(a) Retrospective updating approach



SV = Seen vaccination card
NSV = Not seen vaccination card
□ = classified as unvaccinated
■ = classified as vaccinated
Vac = vaccinated, † = dead.

Retrospective updating approach

In the retrospective updating approach, vaccination status is used as a time-varying variable changing from unvaccinated to vaccinated, on the *exact date of vaccination*. This is a standard statistical approach if vaccination information is collected for all children, regardless of survival status. This approach will introduce *survival bias* if information is missing on vaccinations given since latest visit for children who died. This is illustrated in Figure 1a. For example, if an unvaccinated child is vaccinated between two visits but dies before the last visit, the vaccination card will not be seen and the child continues to be classified as unvaccinated (Figure 1a, child 4). However, if the child survives the vaccination status and is updated on the date of vaccination and the follow-up time, as vaccinated children will be moved to the new vaccination category (Figure 1a, child 3). This latter follow-up time is sometimes referred to as *immortal person-time*, because children are not at risk of dying in the analysis between date of vaccination and date of visit (Rothman & Greenland 1998). Hence, survival bias places immortal person-time in the vaccinated group. Survival bias is a differential misclassification, as the classification as vaccinated depends on the survival of the child.