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.

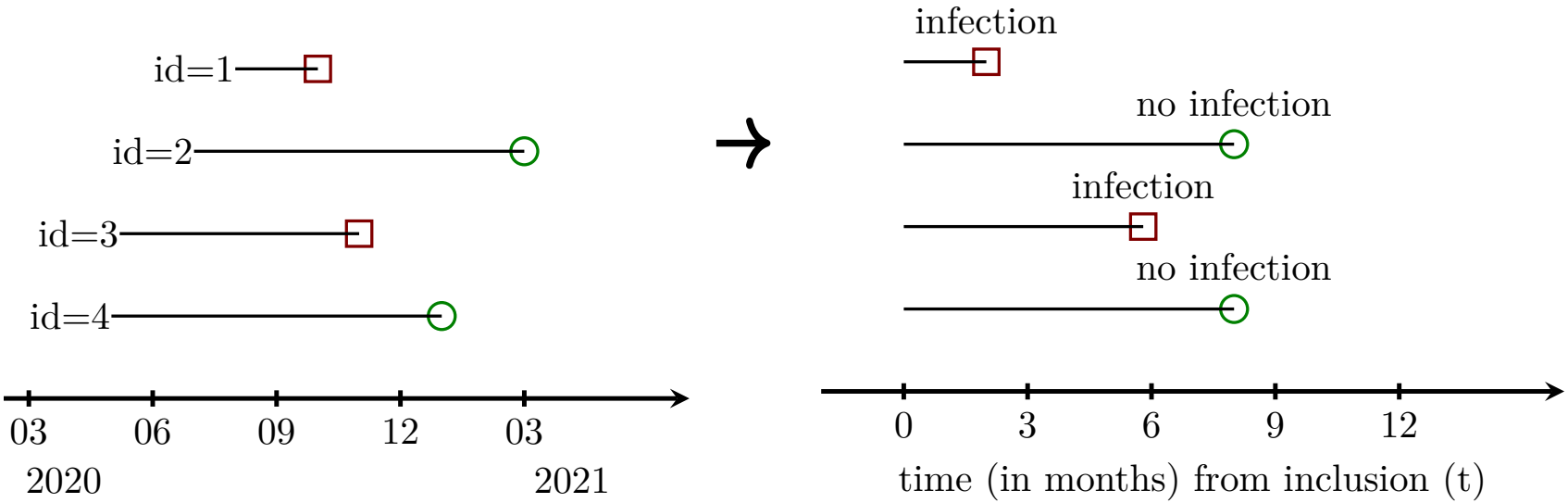
February 28th, 2023 - Brain Drugs WP3

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:

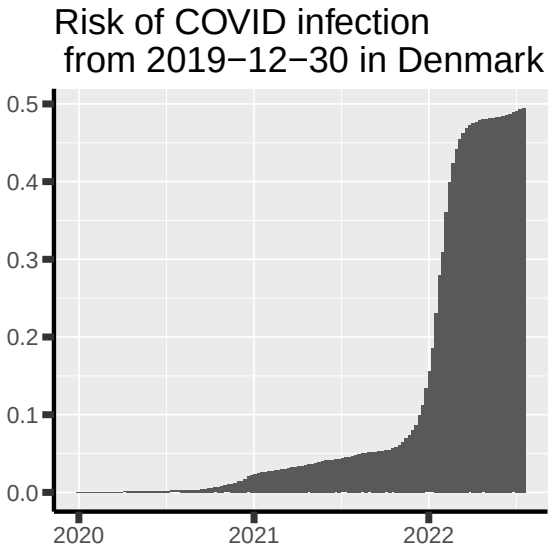
- compare exposed and non-exposed
  with respect to the frequency of the disease/event.
- interpretation and consequences

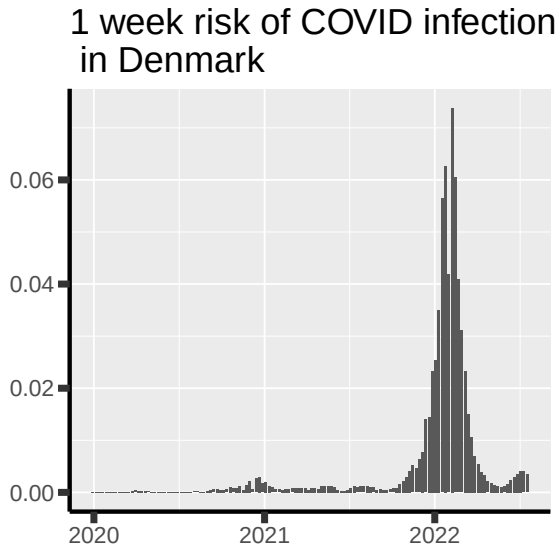
Description of event frequency:

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

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

COVID example (1/2)





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

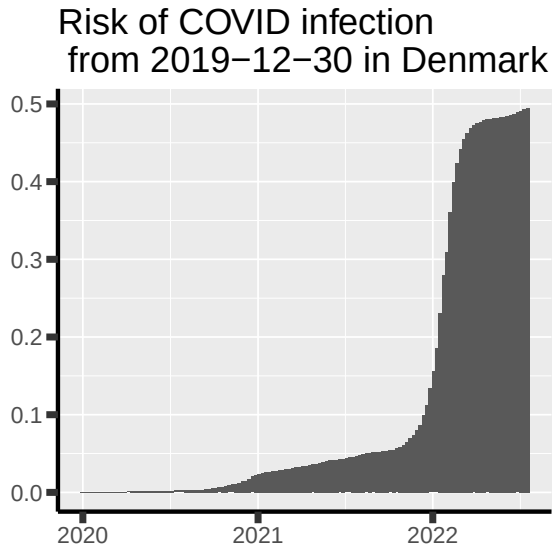
$$r(0;\tau) = \mathbb{P}\left[T \leq \tau, \delta = 1 | T > 0\right] \qquad \in [0,1]$$

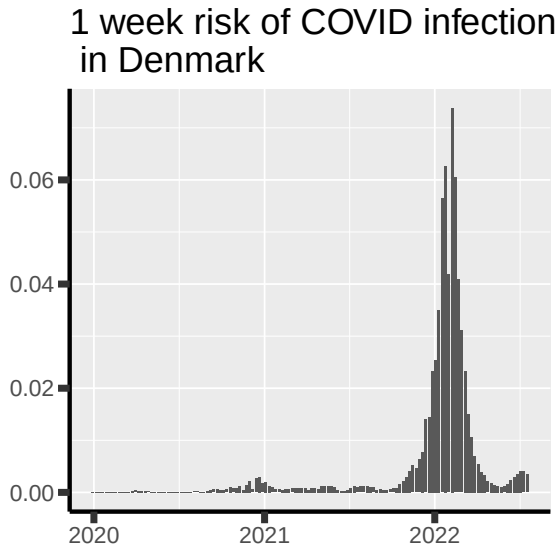
- **incidence rate**: risk of the event divided by at risk time

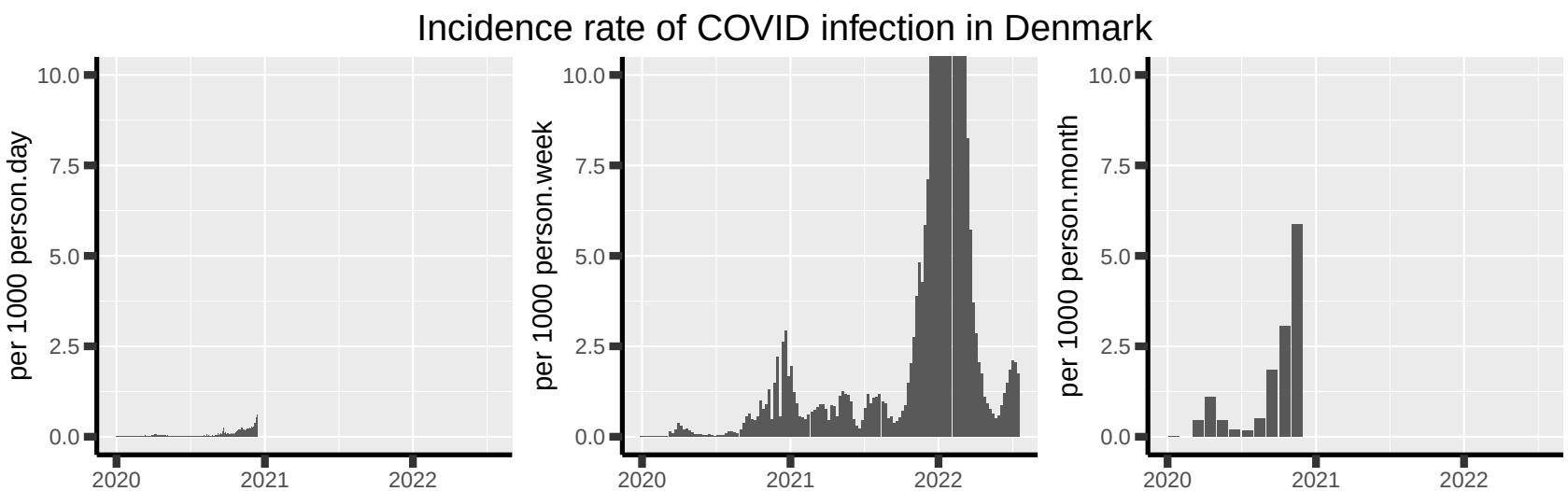
$$\lambda(t;\tau) = \frac{\mathbb{P}\left[T \leq t+\tau, \delta = 1 | T > t\right]}{\tau} \qquad \in [0,+\infty[$$


unit: time^{-1}

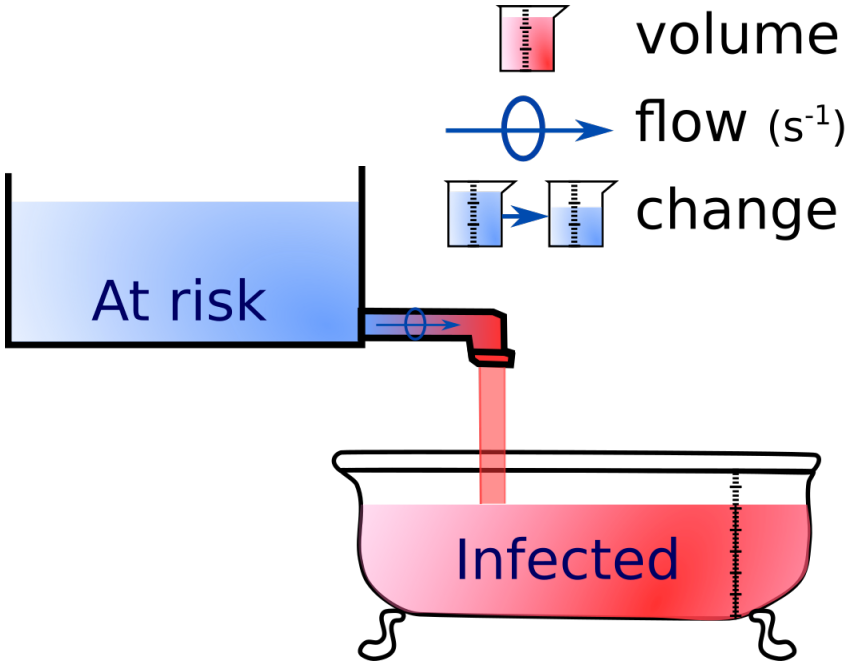
COVID example (2/2)



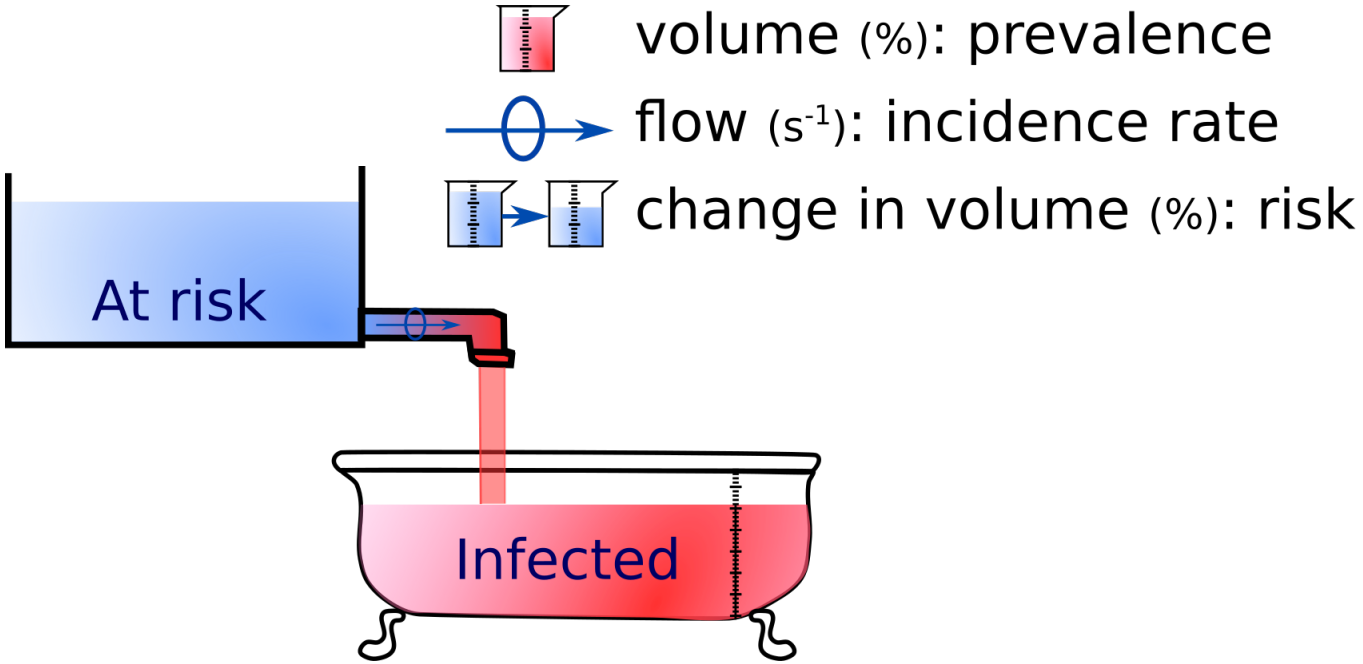




Risk-rate relationship



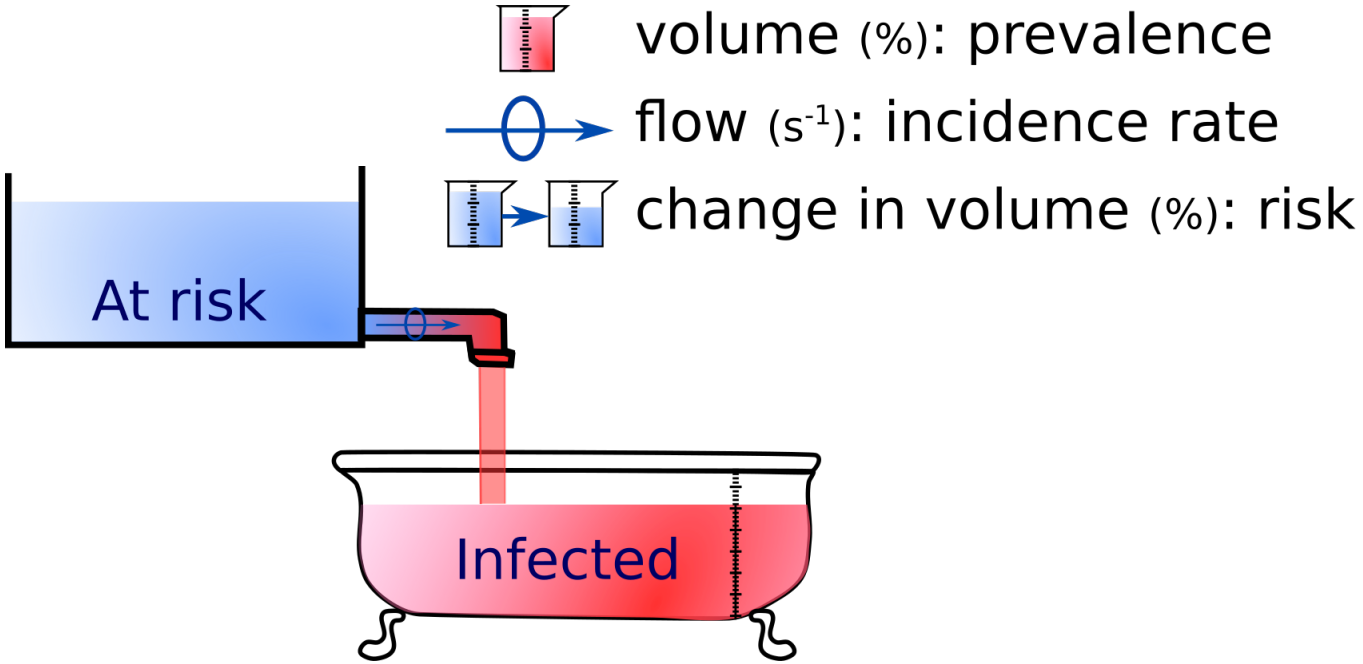
Risk-rate relationship



- *instantenous* rate is also call hazard

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

Risk-rate relationship



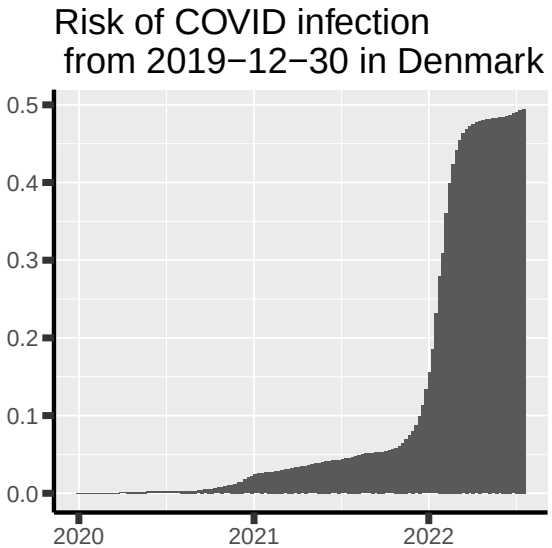
- *instantenous* rate is also call hazard

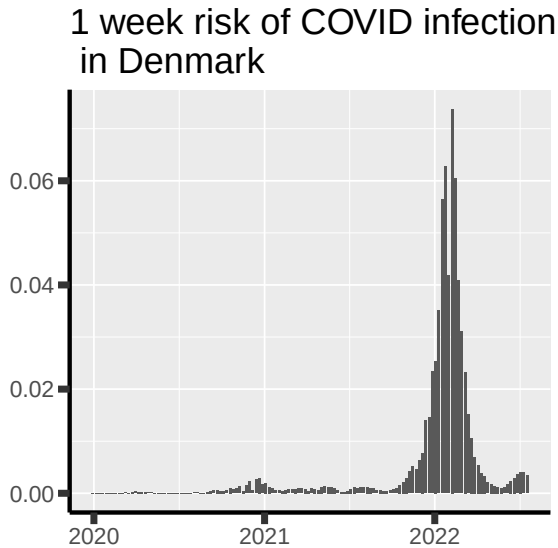
$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}\left[T \leq t + dt, \delta = 1 | T > t\right]}{dt}$$

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

Definition of the parameter of interest

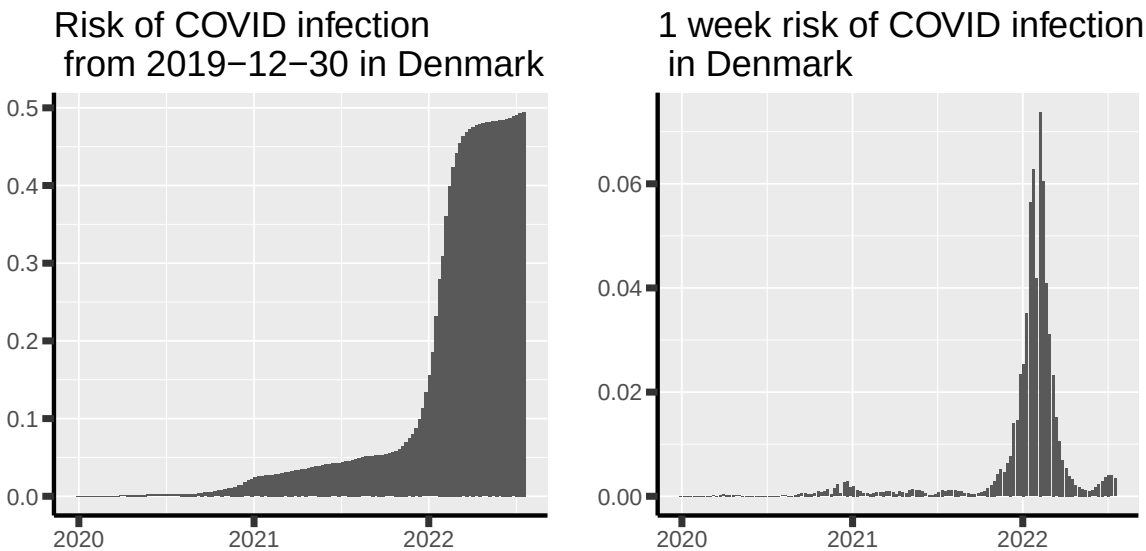
In many medical applications we are interest in the risk





Definition of the parameter of interest

In many medical applications we are interest in the risk

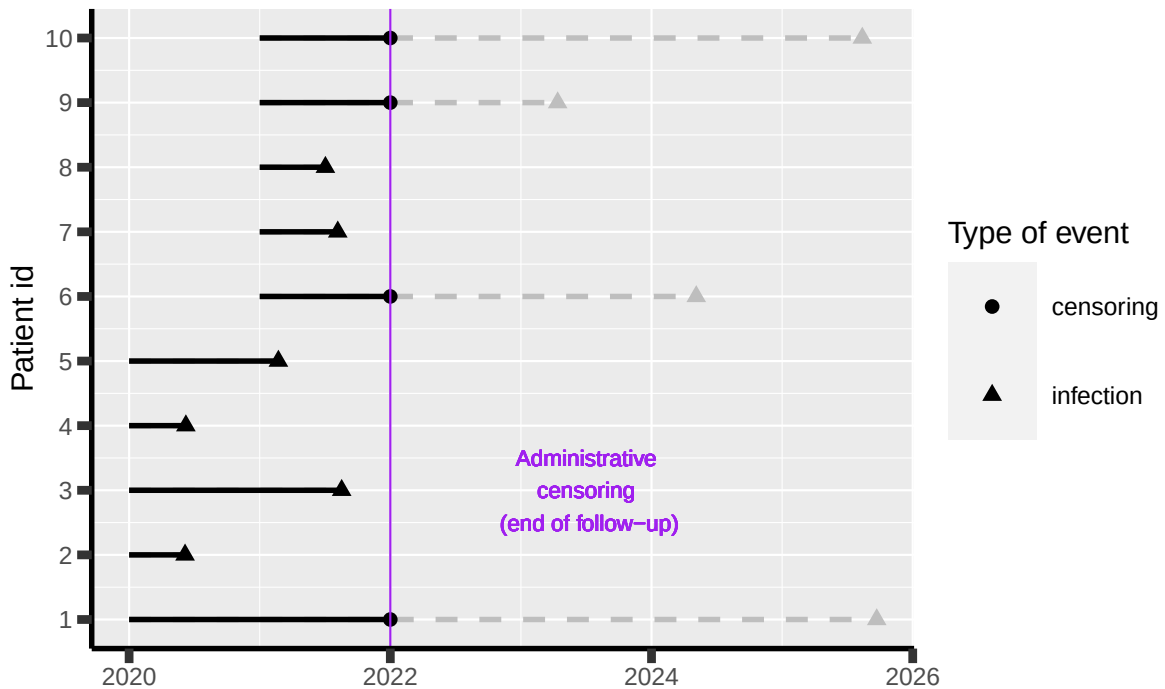


⚠ there is no such thing as "the risk"

- 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

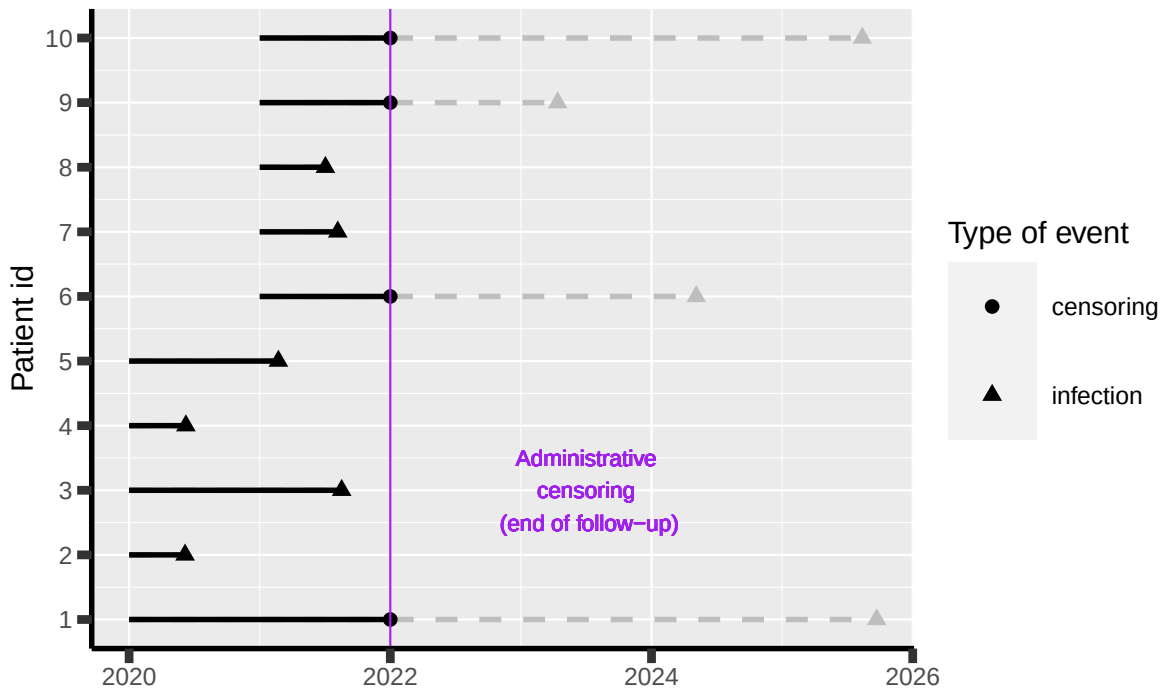
Risk of death between start and end of follow-up: 53.4%



Example

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

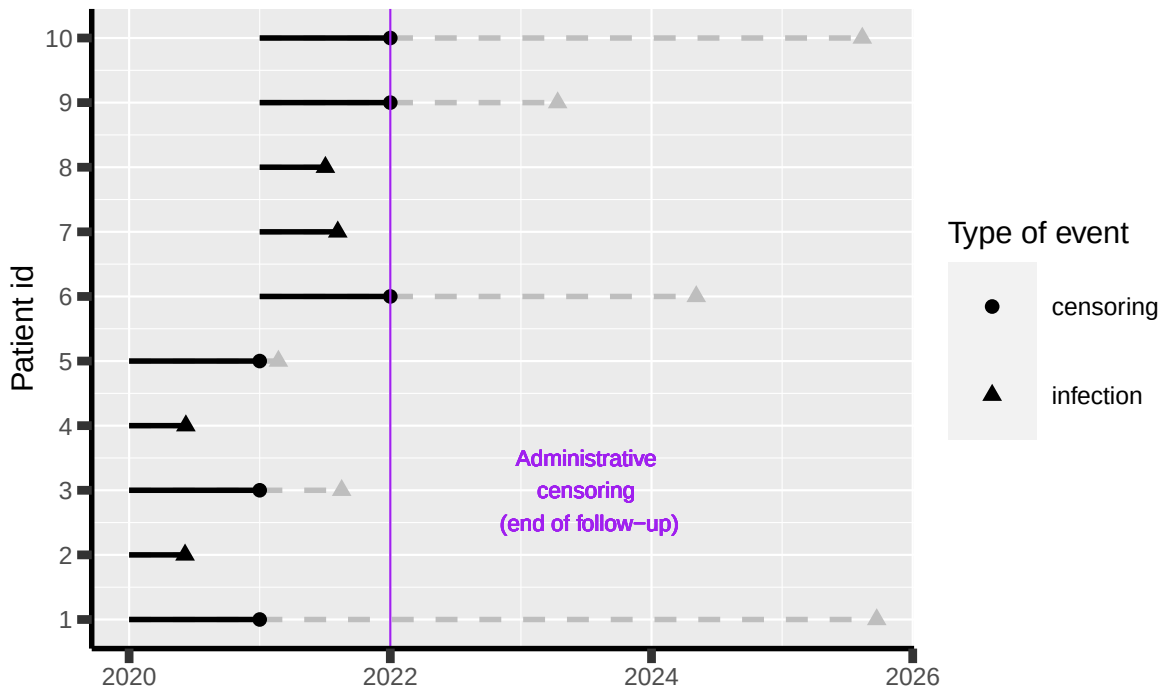
⚠ no clear interpretation! Mix of 5 year risk (42.5%) and 10 year risk (64.2%)



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:

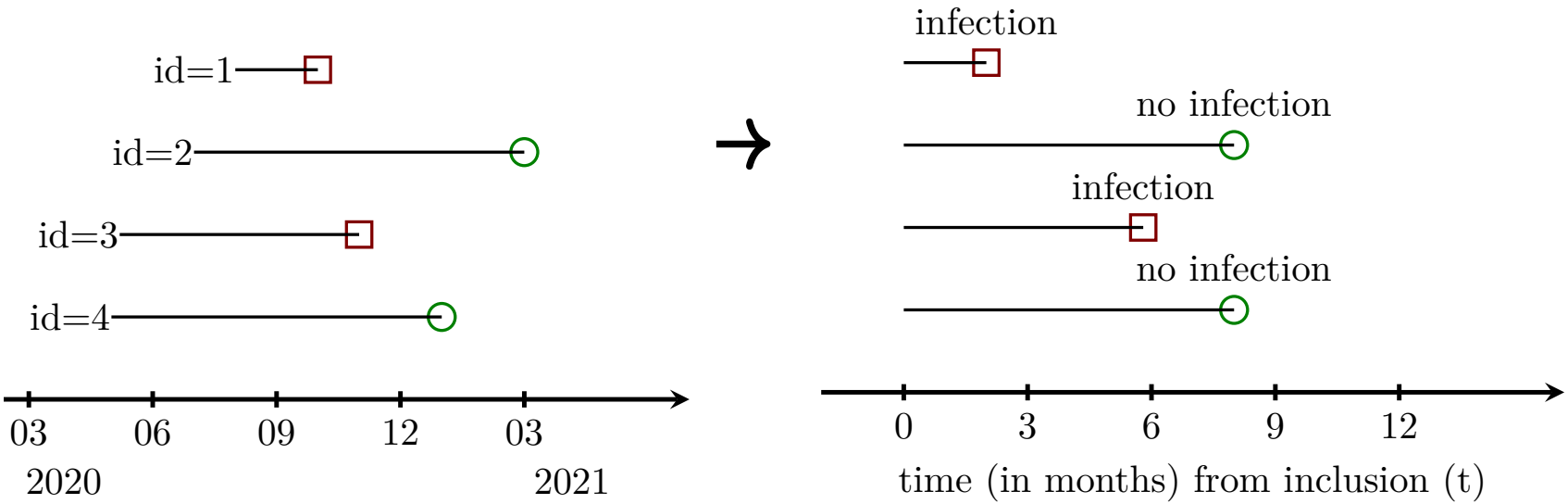
- unambiguously defined and comparable between individuals
- ideally clinically relevant."

"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
- calendar year
- time since treatment initiation.

Time origin - in practice

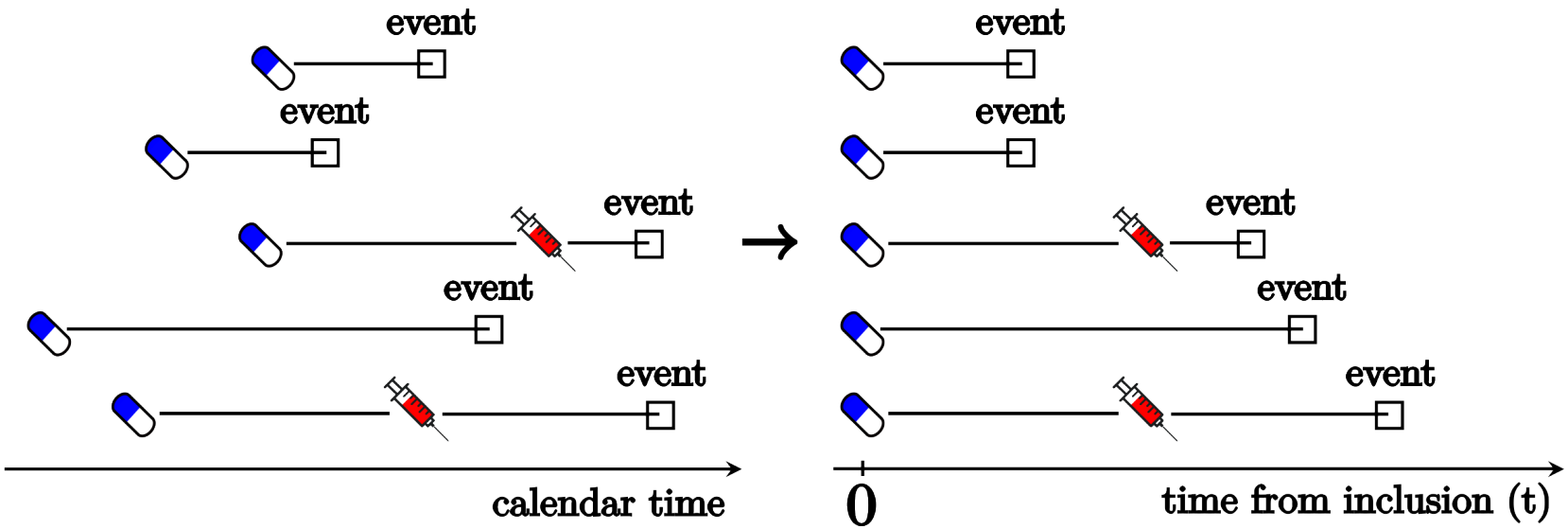


😱 is "time from inclusion" meaningful?

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

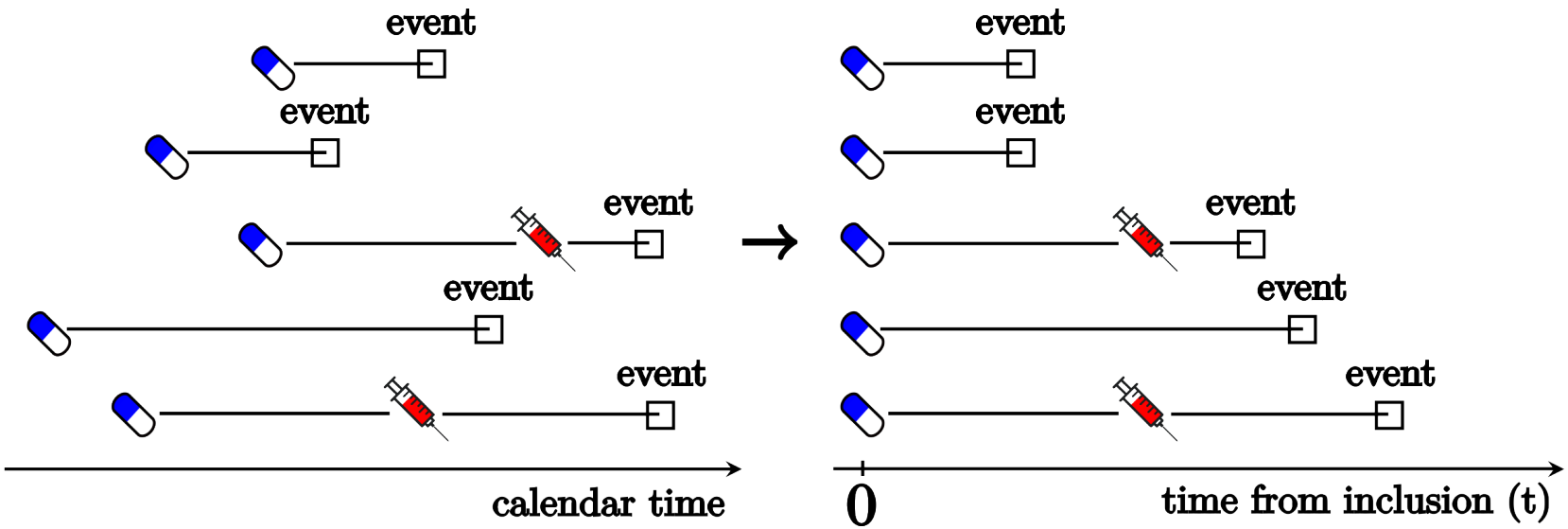
Exposure

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 word

- 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

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

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

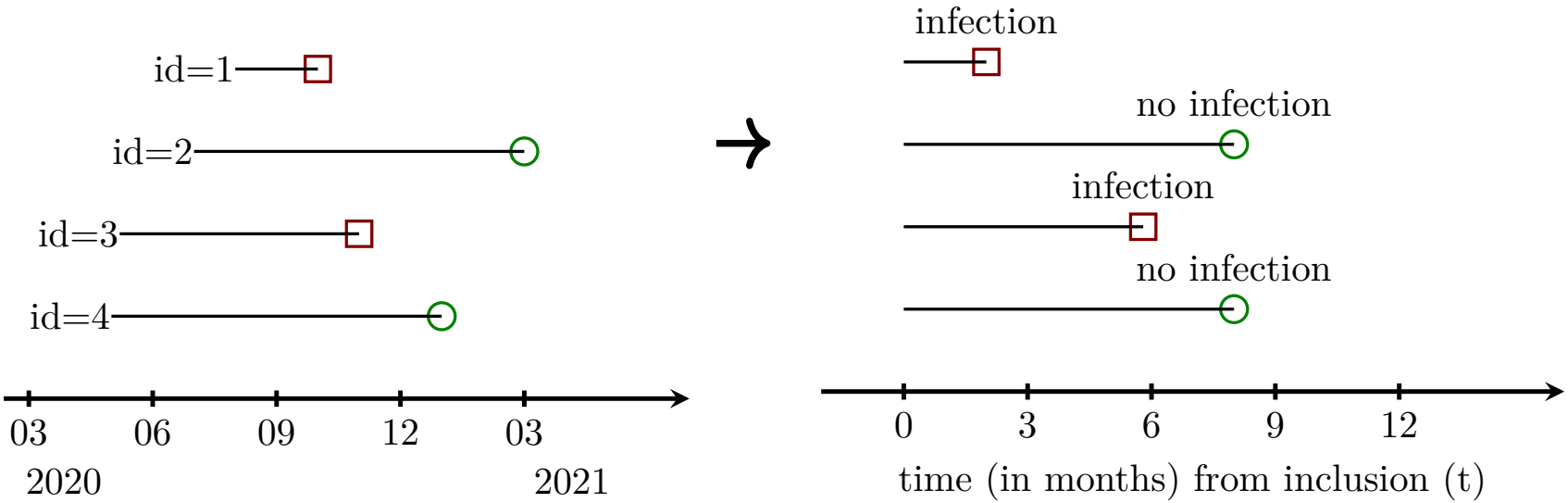
- **incidence rate**: risk of the event divided by at risk time

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

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

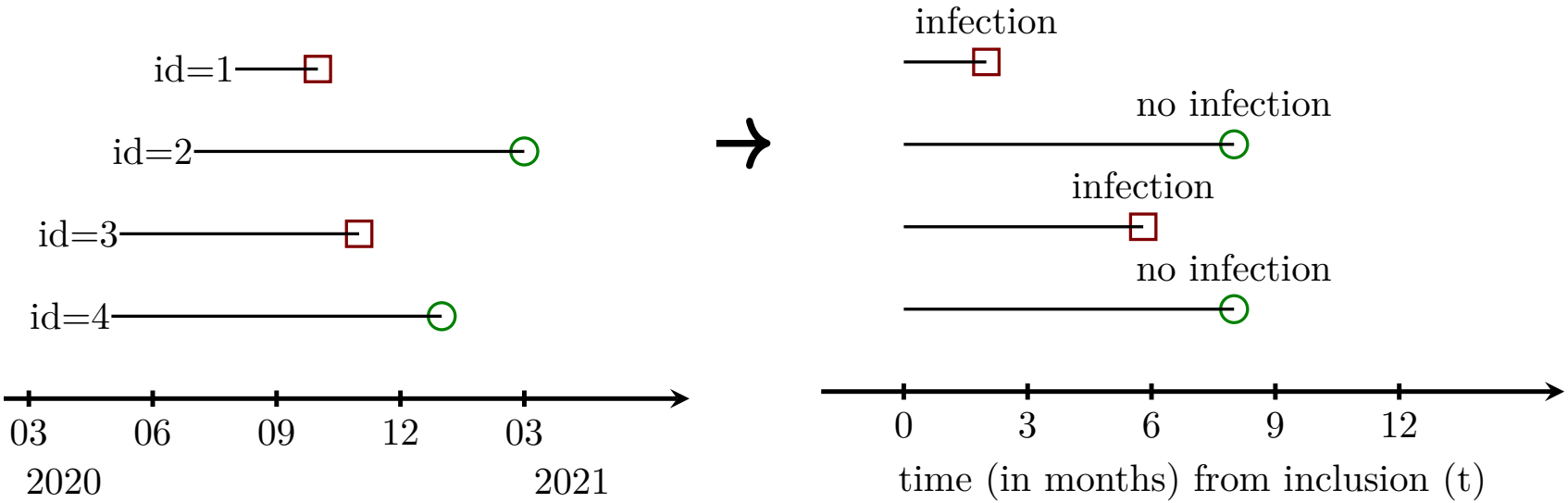
⚠ unit: time^{-1}

Toy example (risk)



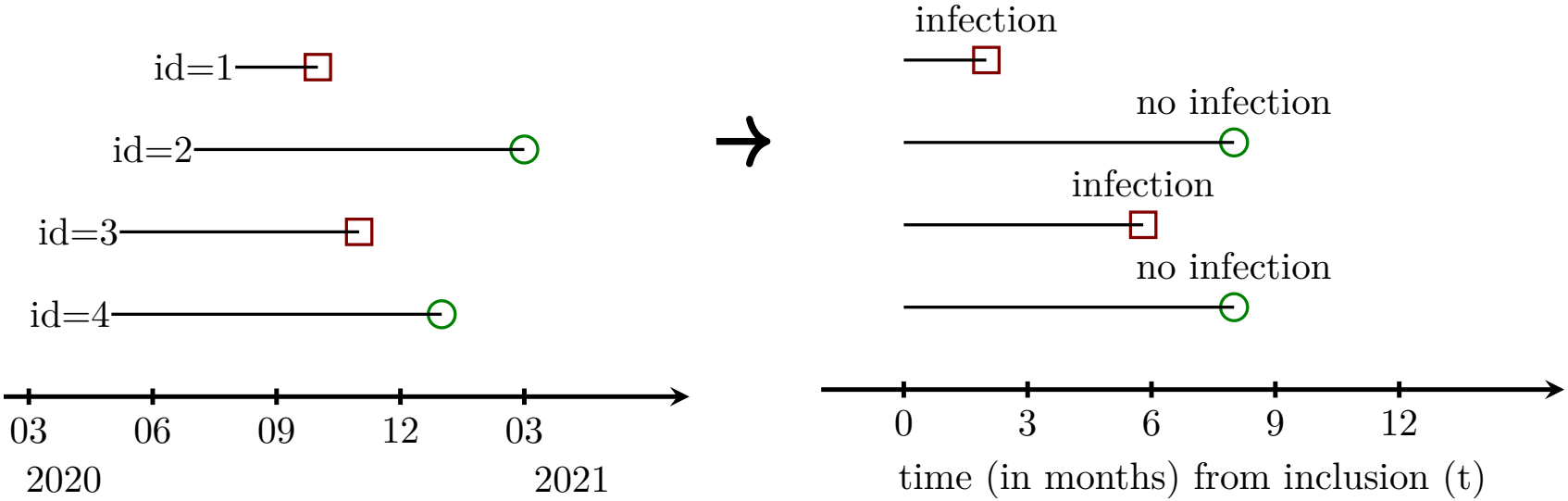
- $\widehat{r}(0) =$ at baseline
- $\widehat{r}(3) =$ after 3 months
- $\widehat{r}(8) =$ after 8 months

Toy example (risk)



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

Toy example (rate)



- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 6$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

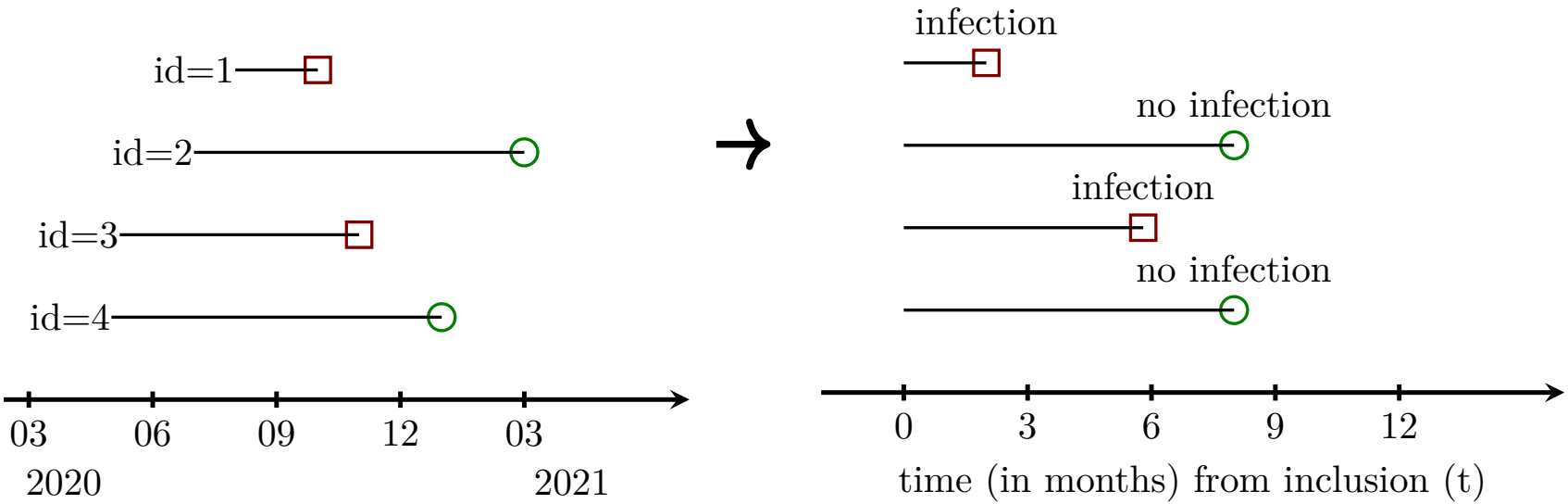
$\widehat{\lambda}_\tau =$

$\approx$ per person-month

$\approx$ per 1000 person-month

$\approx$ per person-year

Toy example (rate)



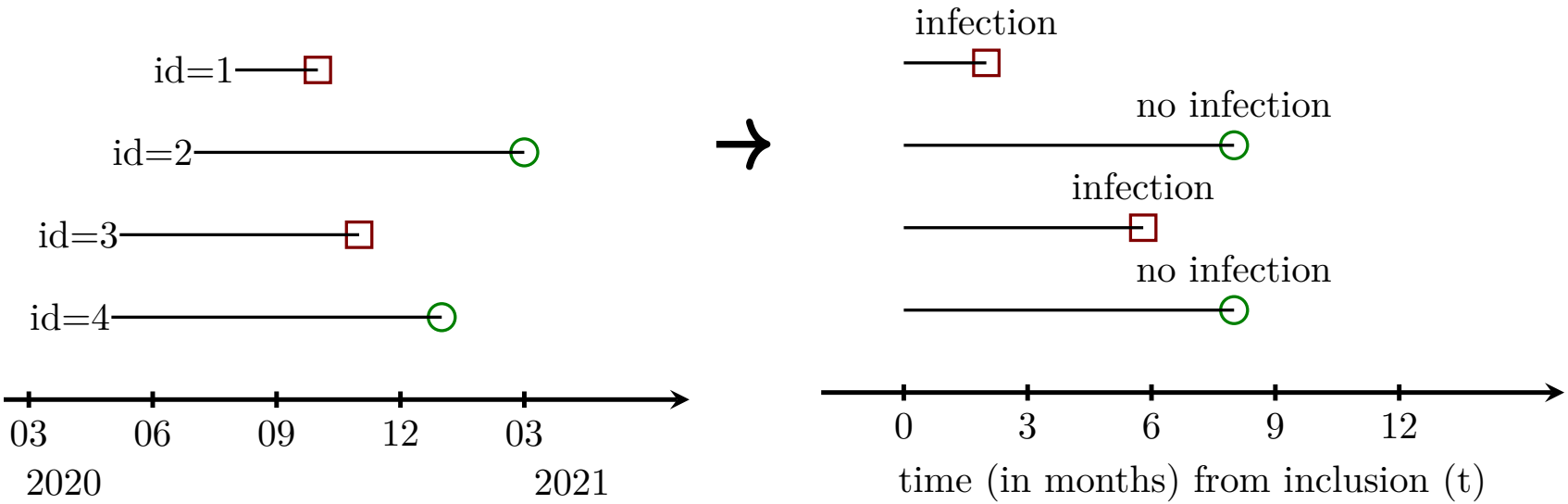
- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 6$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

$$\widehat{\lambda}_\tau = \frac{1+0+1+0}{2+8+6+8} = \frac{2 \text{ new cases}}{24 \text{ person-month}} \approx 0.083 \text{ per person-month}$$

$$\approx 83.33 \text{ per 1000 person-month}$$

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

Toy example (rate)



- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 6$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

$$\widehat{\lambda}_\tau = \frac{1+0+1+0}{2+8+6+8} = \frac{\text{2 new cases}}{\text{24 person-month}} \approx 0.083 \text{ per person-month}$$

$$\approx 83.33 \text{ per 1000 person-month}$$

$$\frac{\text{2 new cases}}{\text{24/12 person-year}} \approx 1 \text{ per person-year}$$

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

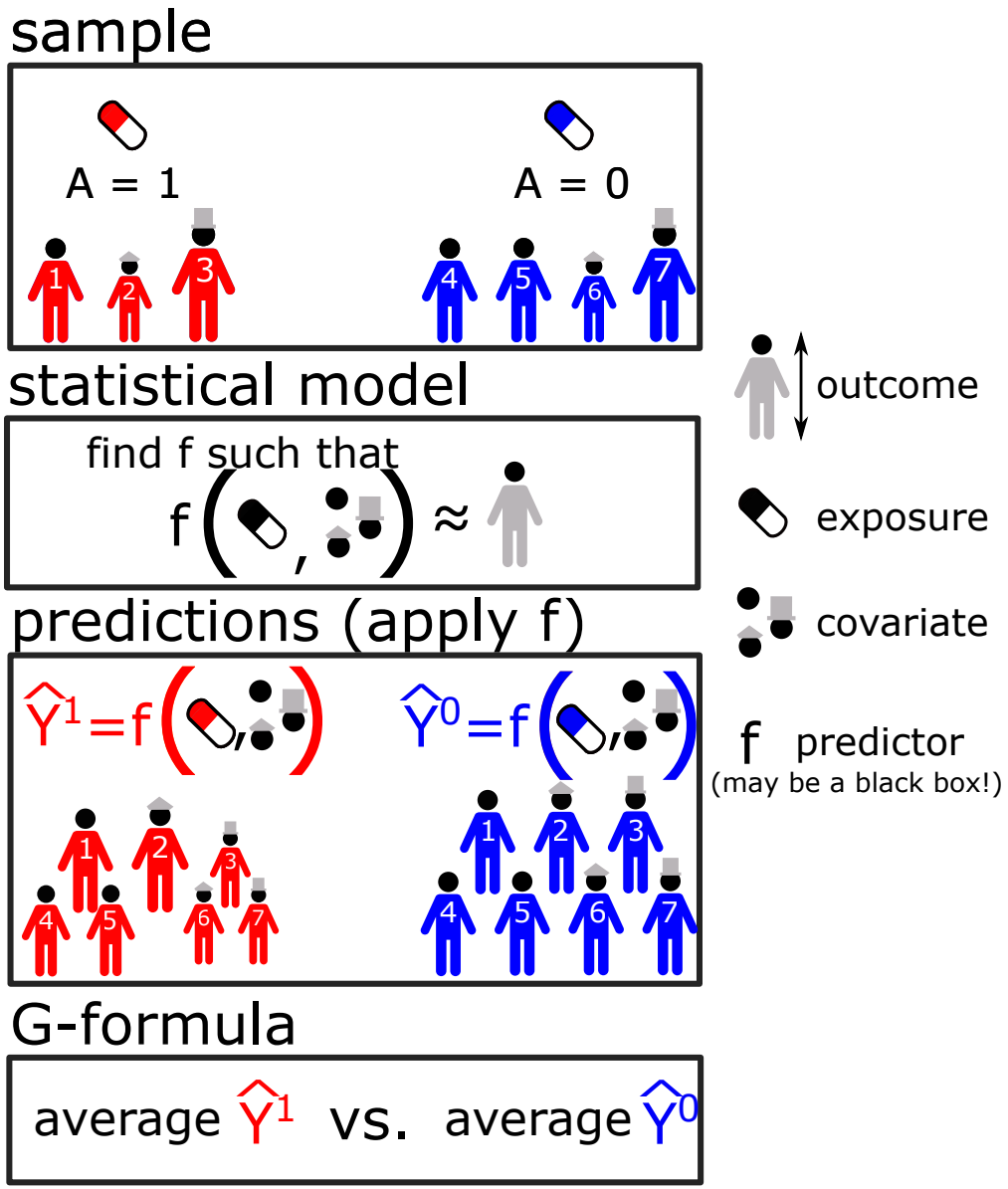
What about heterogeneity in treatment effect?

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 $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3$
- ~~model a constant effect and report this effect~~
- model the age-specific effect and report a standarized effect $\widehat{\Psi} = f\left(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3\right)$

Intuiton behind standardization



Standardization in practice (aka G-formula)

2 equivalent implementations:

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

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

Here for the risk difference:

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

Exercise!

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)
  $\rightarrow$ systematic underestimation of the risk!

⚠ age groups are artificial

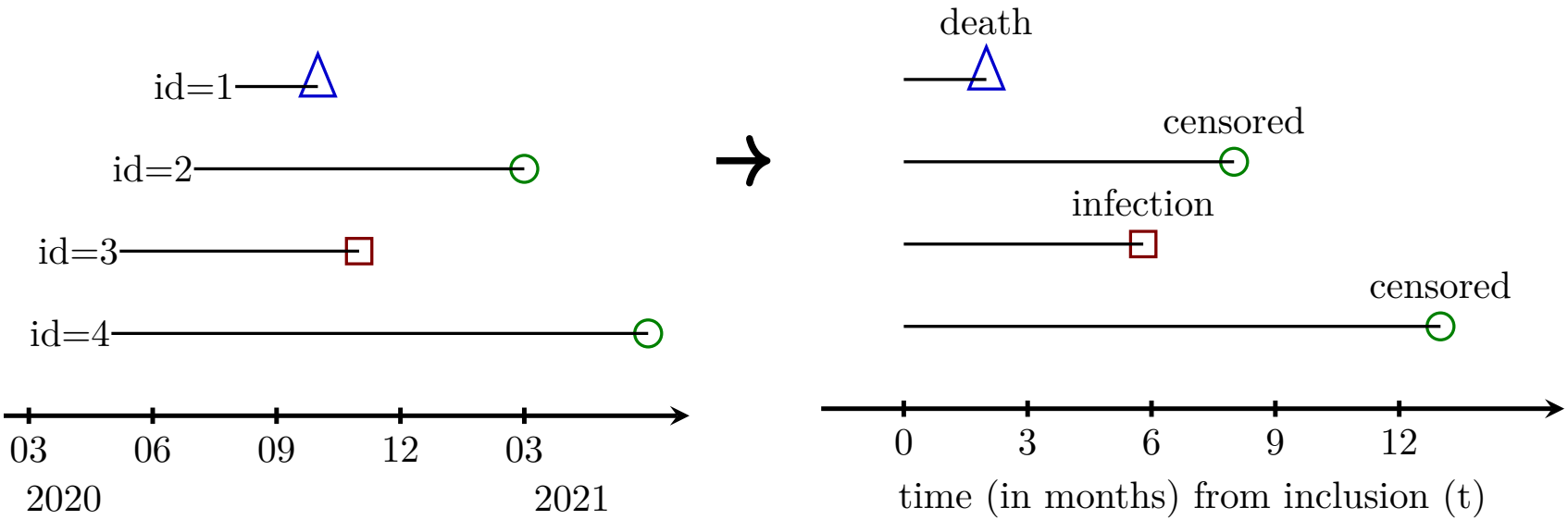
Challenge 1: partially observed outcome

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

- prevent occurence of the event of interest

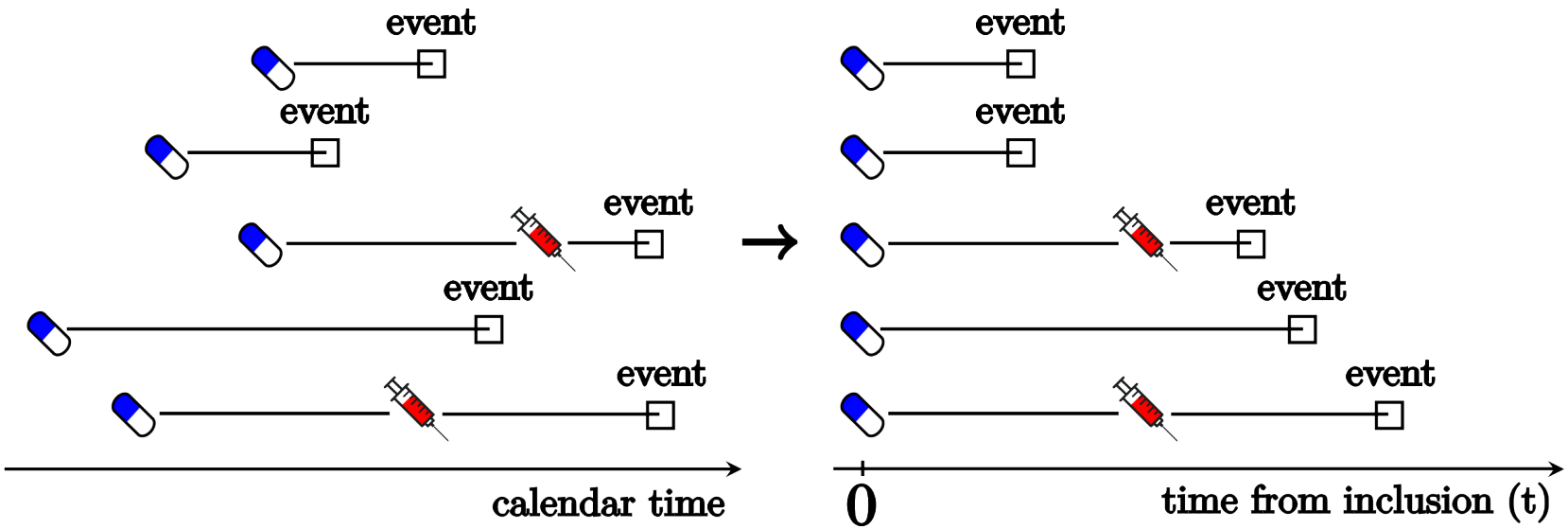
(b) right-censoring:

- event may or may not have occured after last observation



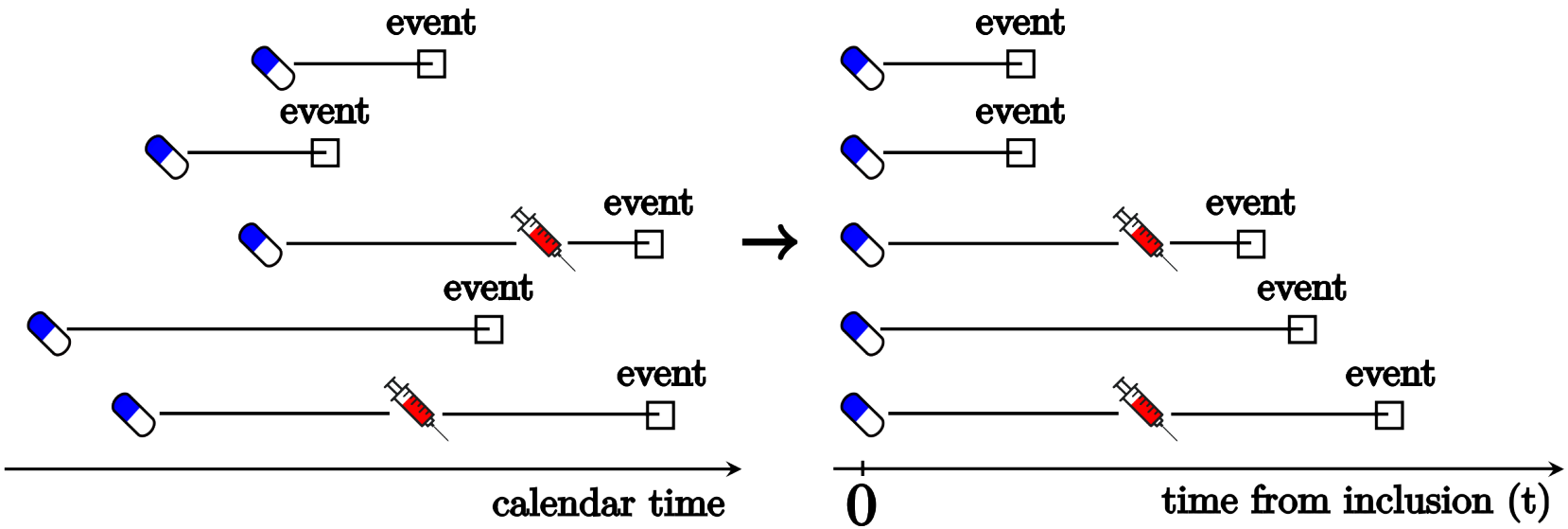
Can we exclude dead/censored patients?
Consider dead patients as free of infection?

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

"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 $\rightarrow$ cannot use 'traditional adjustment'
  (e.g. CD4 counts when studying HIV treatments)

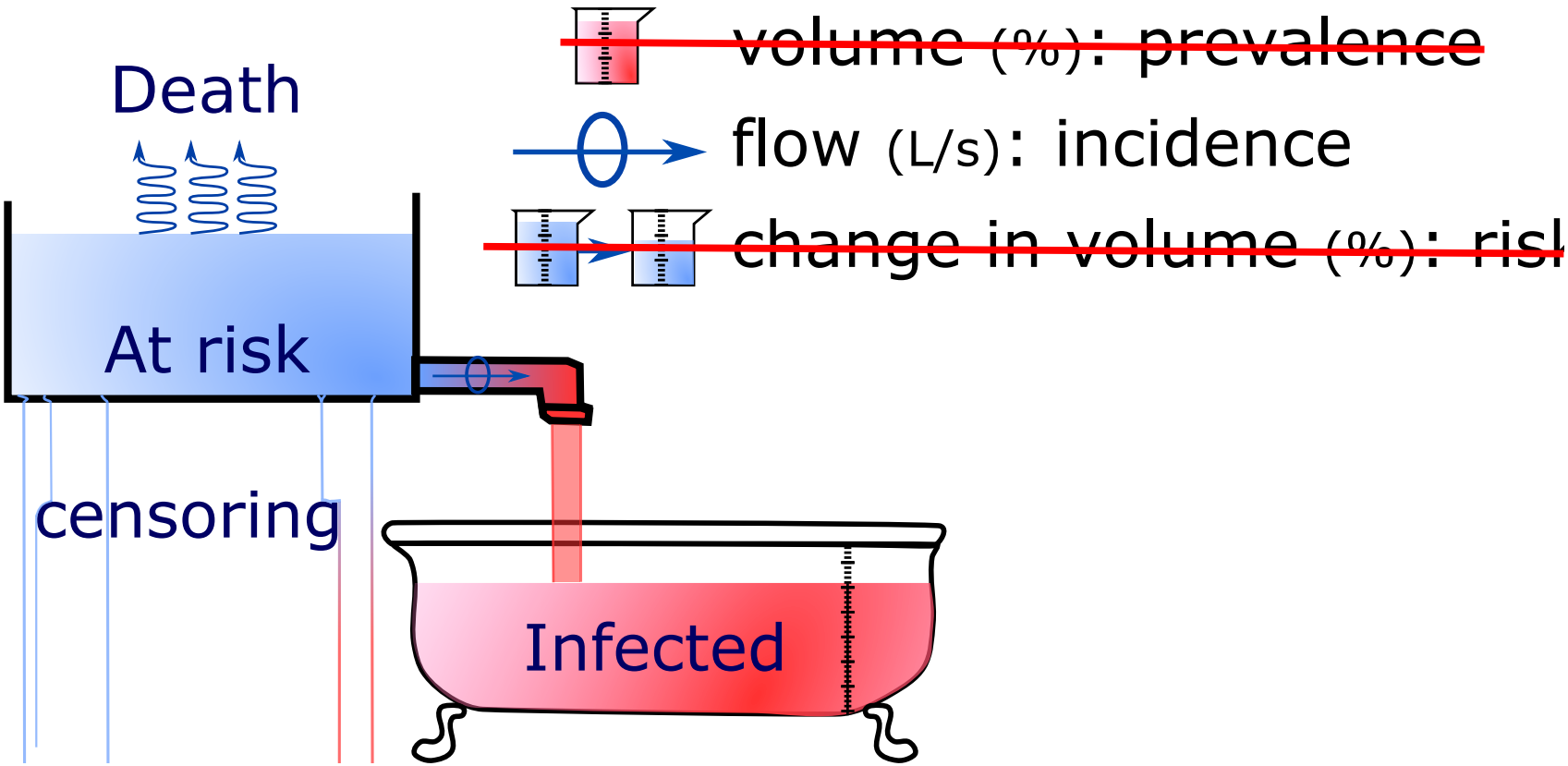
Complex exposure:

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

Big picture

Because of complications we will (often) model the incidence

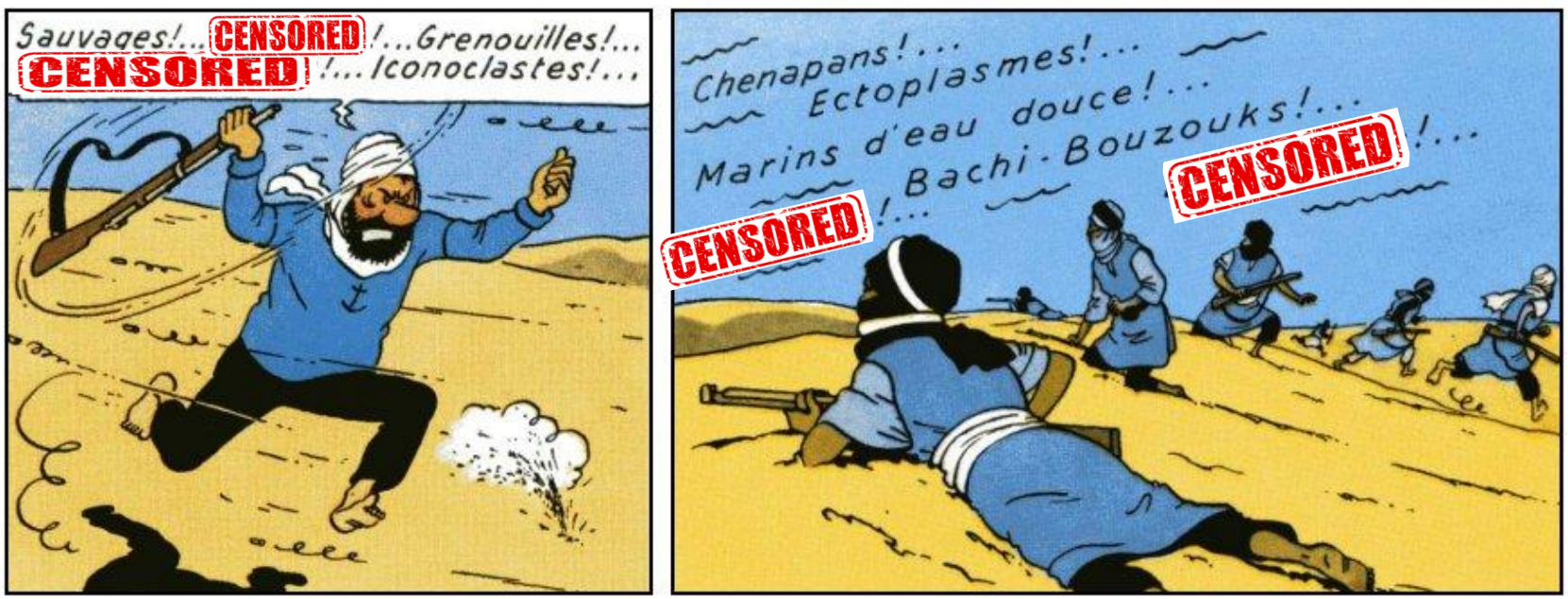
- and then deduce the risk



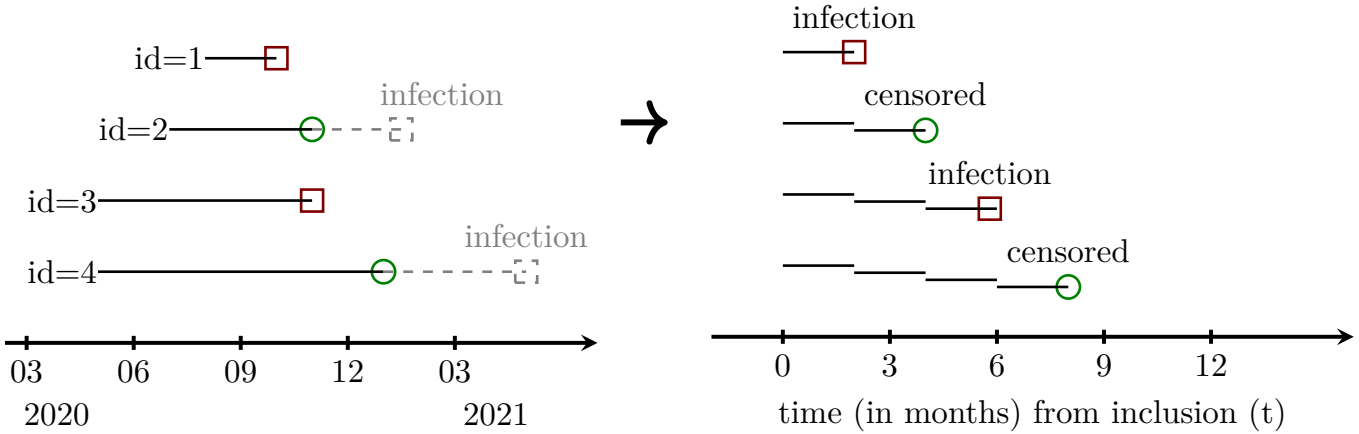
⚠ do not loose track of what you want because of a detour!

Handling censoring

- From rates to the Kaplan Meier estimator
- Kaplan Meier estimator as a weighting approach
  - independent censoring assumption



Toy example (risk under censoring)



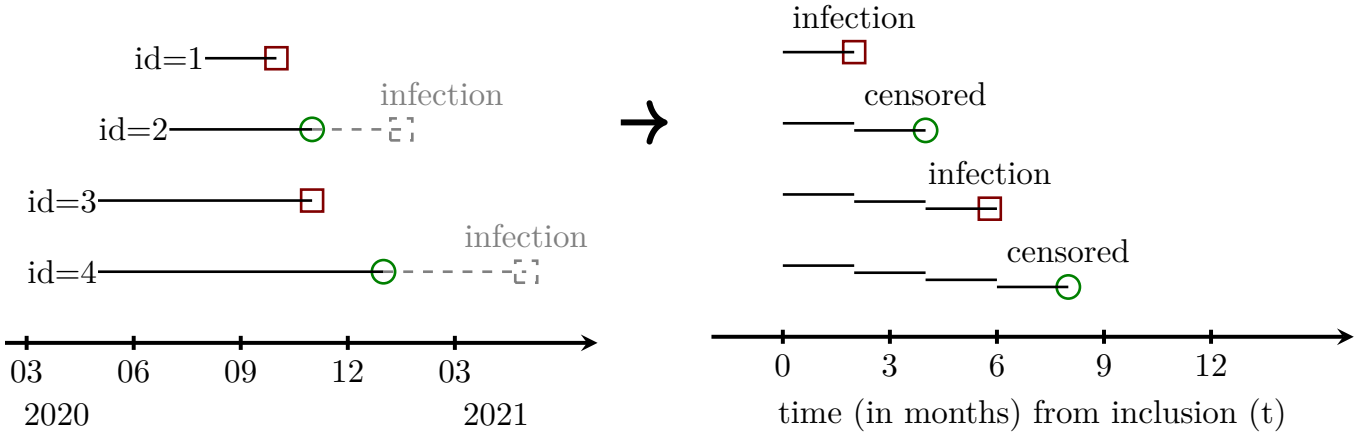
Risk after 8 months:

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

Incidence:

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

Toy example (risk under censoring)



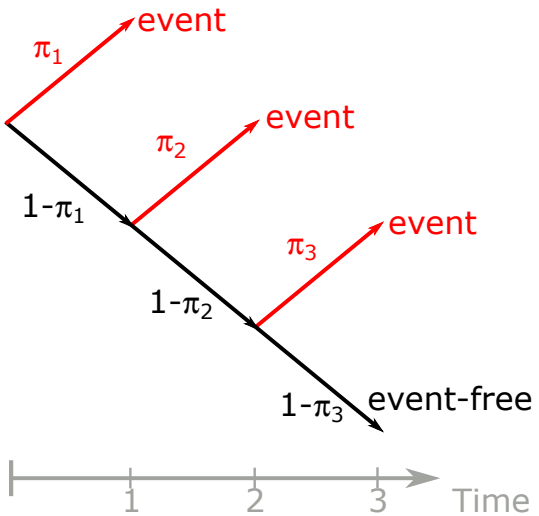
Risk after 8 months:

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

Incidence:

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

Binary probability models



Survival (probability of not getting the event)

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

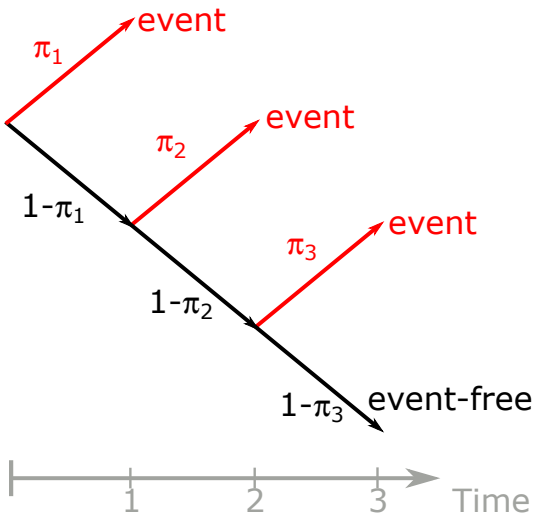
$$=$$

Risk (probability of getting the event)

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

$$=$$

Binary probability models



Survival (probability of not getting the event)

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

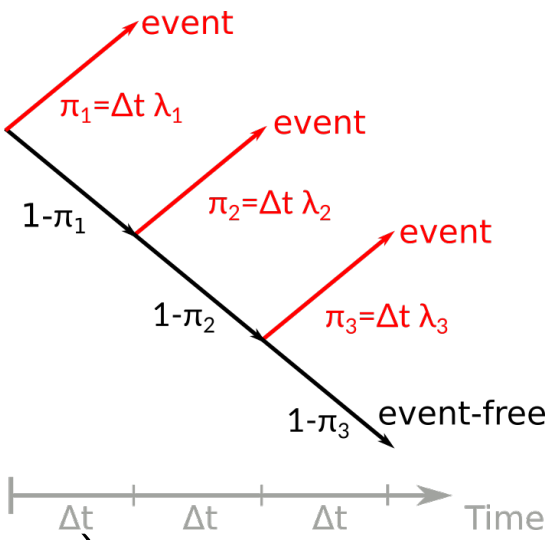
Risk (probability of getting the event)

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

Binary probability models

Assuming piecewise constant hazard:

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



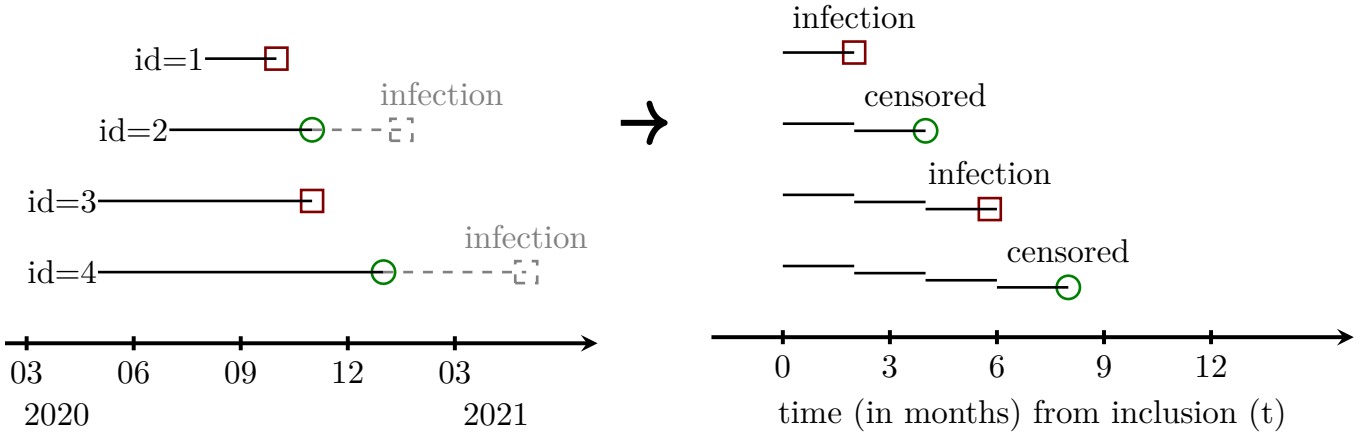
Survival (probability of not getting the event)

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

Risk (probability of getting the event)

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

Toy example (risk under censoring)



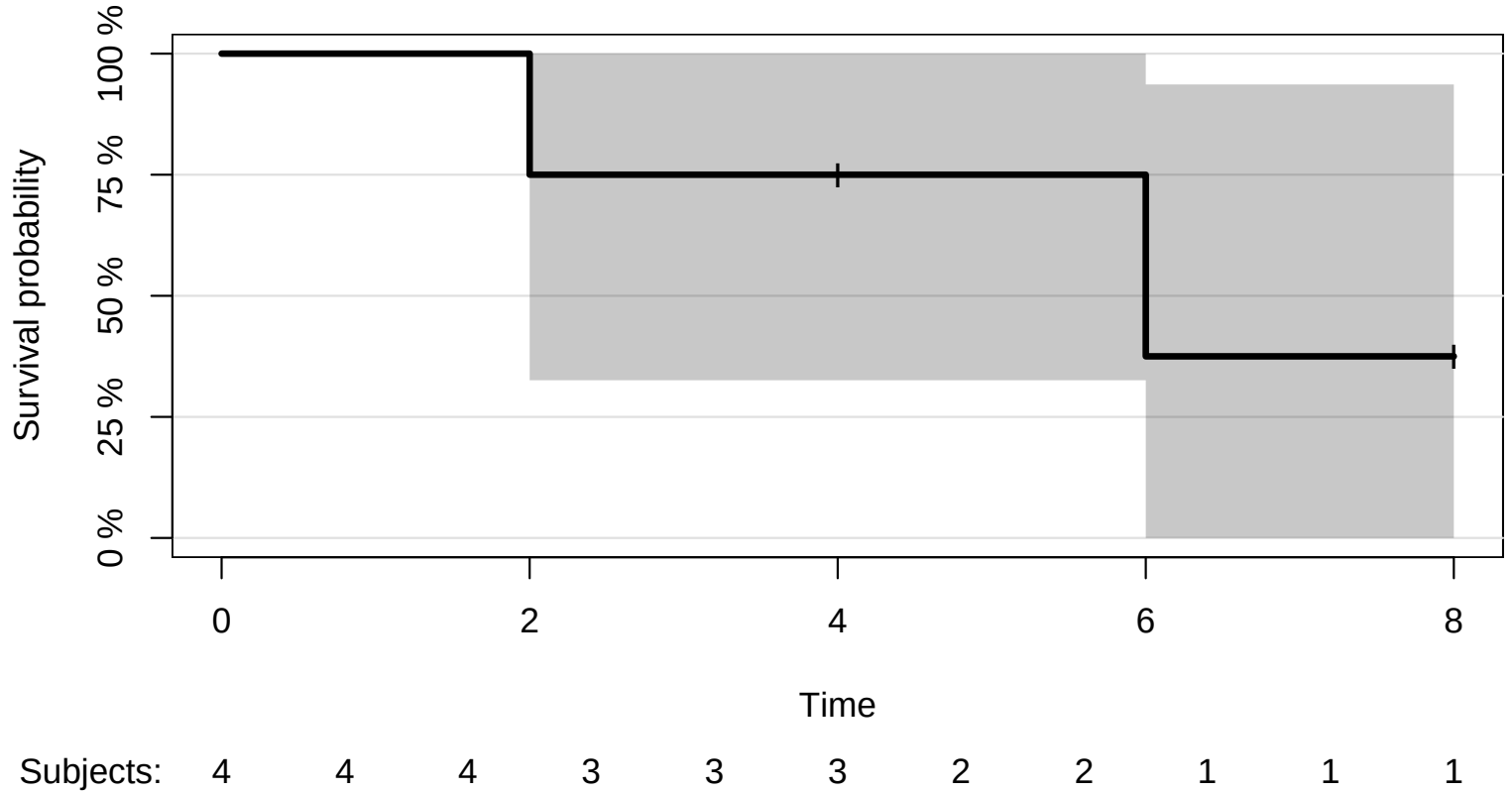
Risk after 8 months:

- $\widetilde{r}(8) = 0.5$ or 0.75
- $\widehat{r}(8) = 1 - (1 - \widehat{\lambda}_1 \Delta t)(1 - \widehat{\lambda}_2 \Delta t)(1 - \widehat{\lambda}_3 \Delta t)(1 - \widehat{\lambda}_4 \Delta t)$
  $= 1 - (1 - 1/8 * 2) * 1 * (1 - 1/4 * 2) * 1 = 0.625$

Incidence:

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

Kaplan Meier in R




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

Exercise!

File *exercise-workshopEpi.R* (line 99-156)

Generate, visualize, and analyse the toy example

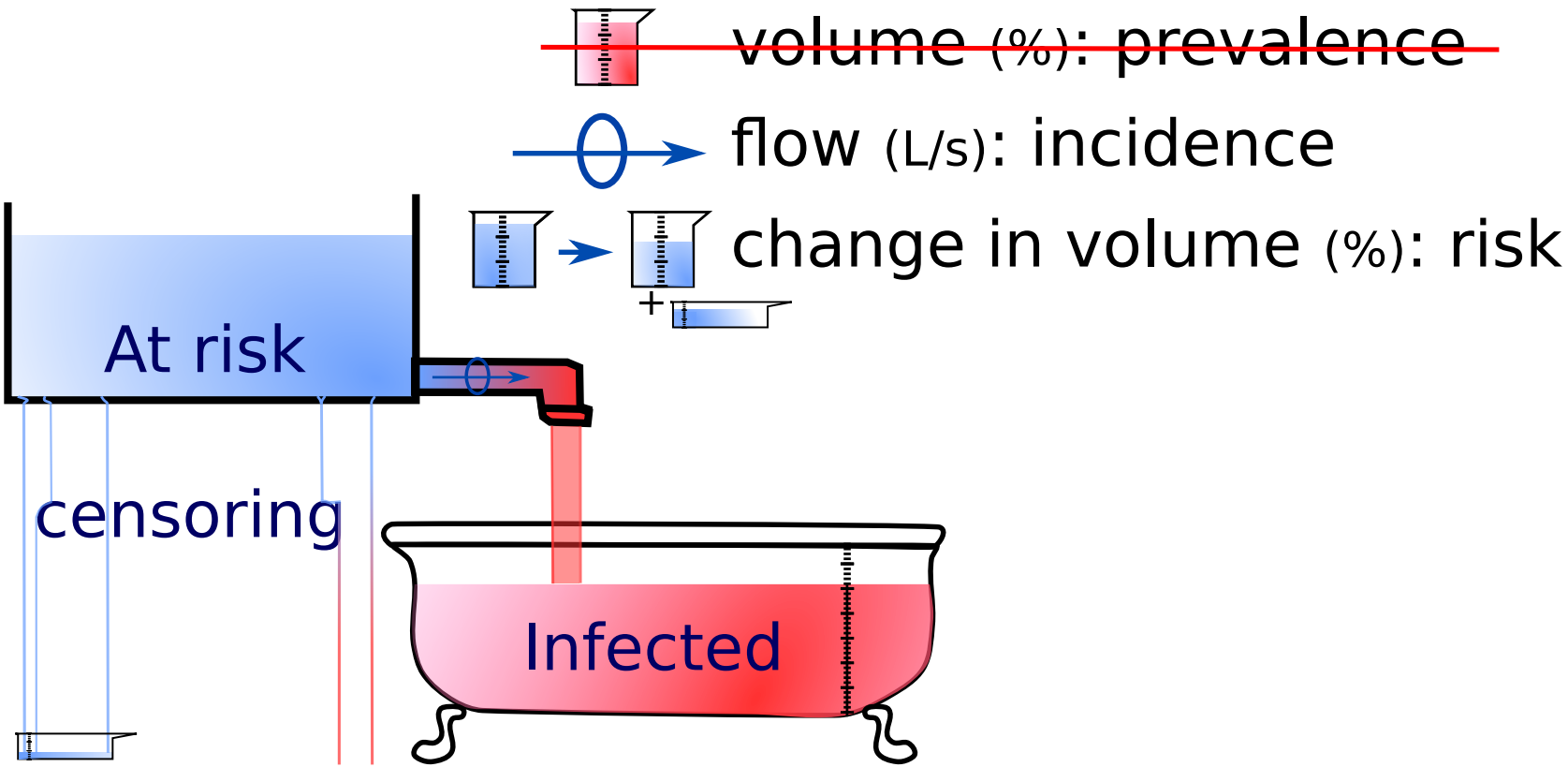
- computing the rate and deducing the risks
- using the Kaplan-Meier estimator to estimate the risks

Re-analyze the data from the Bissau study:

- estimate the risks, accounting for right-censoring
- compare the risks with those 'ignoring censoring'?

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:

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

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:

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

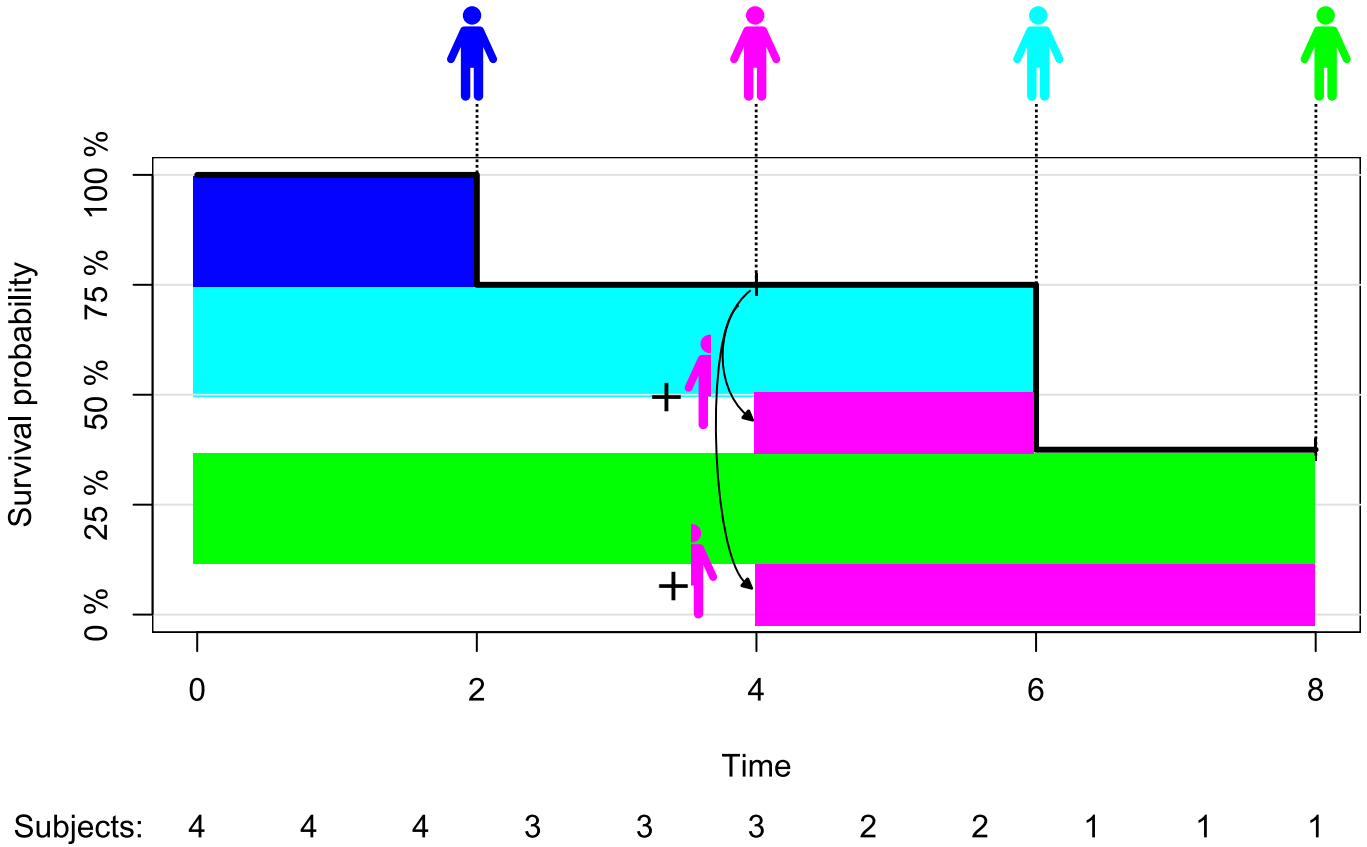
where T_i is the time to event for individual i.

We now also consider C_i, the time to censoring.
$\delta_i \in \{0, 1\}$ indicates whether censoring or event is observed.

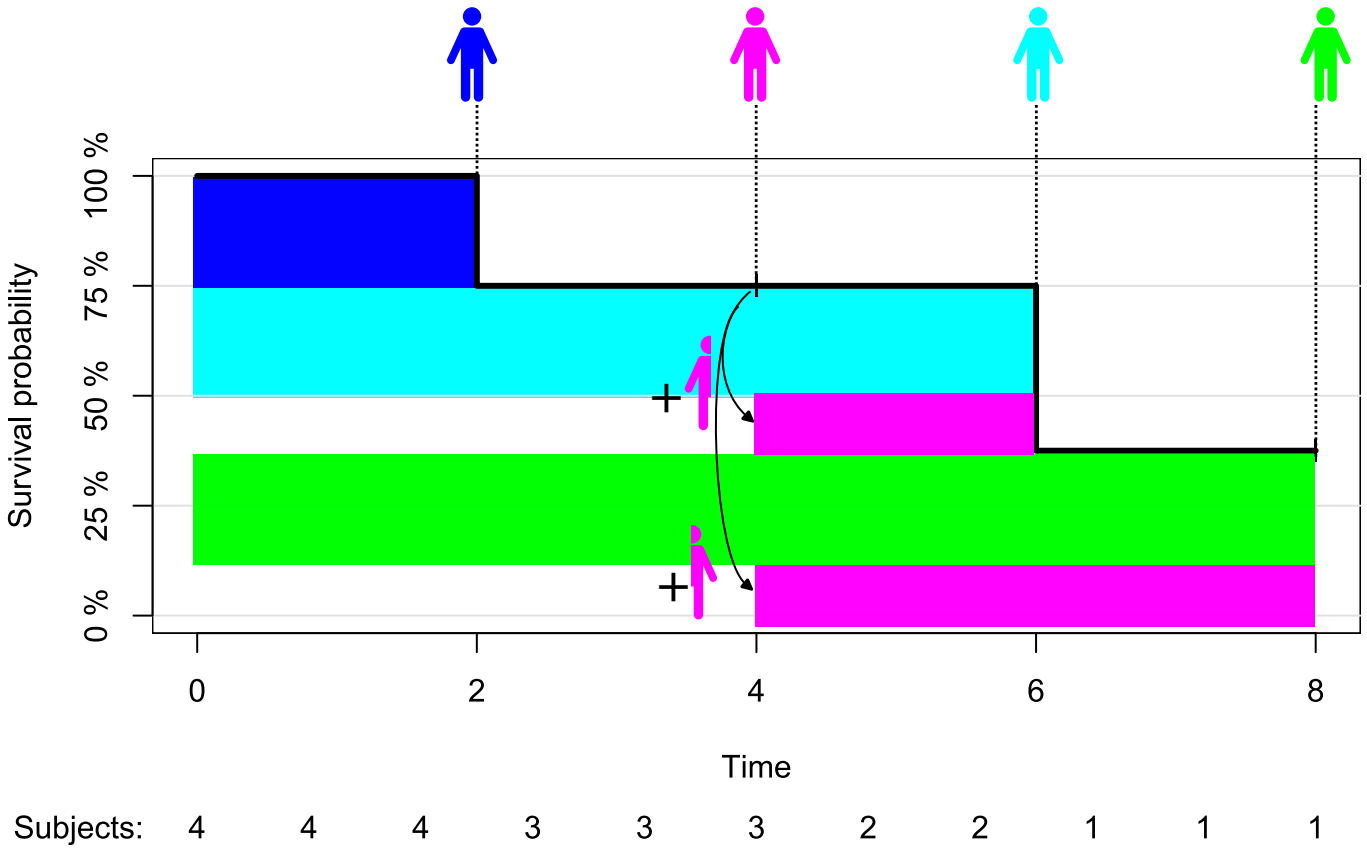
- censored observations at time t will not contribute
- uncensored observations at time t will contribute, weighted by the inverse of their probability to be observed.

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

Efron redistribution algorithm



Efron redistribution algorithm



- patients who stay are **representative** of those who drop-out
- we evaluate the survival effect **had nobody been censored**! (same for the risk or treatment effect)

Independent censoring assumption

The censoring status of a currently event free patient should not be informative of his risk of infection at any later timepoint.

✔ 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)

Independent censoring assumption

The censoring status of a currently event free patient should not be informative of his risk of infection at any later timepoint.

✔ 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)

- how could this assumption be violated in the bissau study 

Exercise!

File *exercise-workshopEpi.R* (line 158-188)

Run the code analyzing the toy example with IPCW

- compare to the Kaplan Meier approach

Run the code analyzing the bissau study with IPCW

- compare to the 'ignoring censoring' approach

Summary

Two (main) approaches for handling right-censoring:

- modeling the rate and deducing the risk
  - ✔ less modeling (no censoring model)
  - ✔ traditional approach
  - ⚠ modeling on the rate instead of risk scale

- modeling the censoring process to re-weight the observations when modeling the risk (IPCW)
  - ✔ modeling on the risk scale
  - ✘ less efficient estimator (but improvements exist)

Key assumptions:

- population of interest: had nobody been censored
- independent censoring assumption

Handling competing risks

- absolute risk / cumulative incidence function
  - Aalen Johansen (AJ) estimator

Competing risks

Patient may experience events:

- preventing the event of interest (e.g. death)
- making the event of interest no more relevant (e.g. bipolar disorder when studying depression)

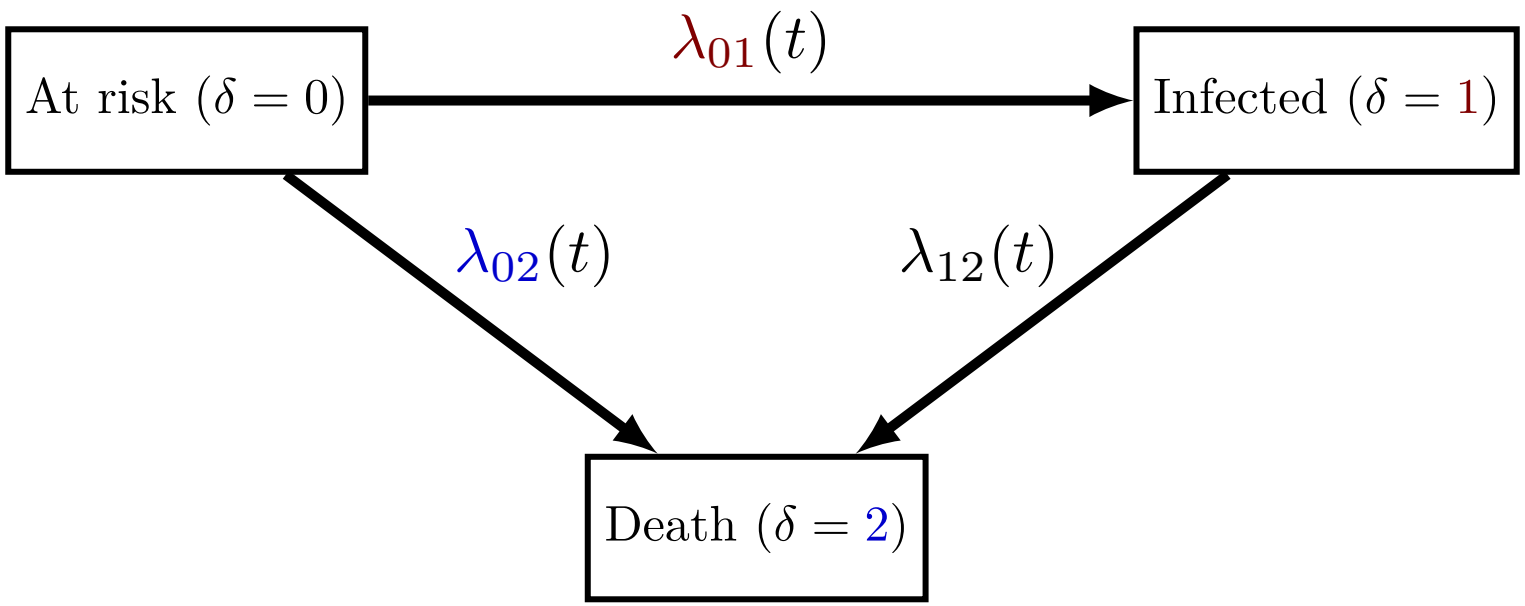
$\rightarrow$ likelihood increases with follow-up time

Competing risks

Patient may experience events:

- preventing the event of interest (e.g. death)
- making the event of interest no more relevant (e.g. bipolar disorder when studying depression)

$\rightarrow$ likelihood increases with follow-up time



Exercise!

`Melanoma`: Data of the survival of 205 patients with malignant melanoma (skin cancer) after surgery betwen 1962 and 1977

we will work on an artificial dataset without censoring `Melanoma2`

File *exercise-workshopEpi.R* (line 190-217)

Compute the risk of death, cancer related death, death due to other causes as a proportion of events

- how does it compare to using Kaplan-Meier?
- which approach seems the most reasonnable?

Classical mistakes (Andersen et al., 2012)

1. Treating competing events as censorings:

Classical mistakes (Andersen et al., 2012)

1. Treating competing events as censorings:

- is conceptually wrong: risk had nobody been censored **or died**!
  $\rightarrow$ violate principle 3!
  $\rightarrow$ do not use Kaplan Meier!
- gives wrong results: upwards biased estimate of the risk
  (since the event is no more prevented by death)

Re-defining the risk (1/2)

Consider equally spaced timepoints $t_1 = 1,\ t_2 = 2,\ \ldots,\ t_k = t$

$$
\begin{aligned}
r_1(t) &= \mathbb{P}\left[T \leq t, \delta = 1\right] \\
&= \mathbb{P}\left[T = 1, \delta = 1\right] + \mathbb{P}\left[1 < T \leq 2, \delta = 1\right] + \ldots \\
&= \mathbb{P}\left[T = 1, \delta = 1\right] + \mathbb{P}\left[T = 2, \delta = 1 | T > 1\right] \mathbb{P}\left[T \geq 1\right] + \ldots \\
&= \lambda_{01}(1) + \lambda_{01}(2) S(1) + \ldots \\
&= \int_{s=0}^{t} \lambda_{01}(s) S(s-) ds
\end{aligned}
$$

Re-defining the risk (1/2)

Consider equally spaced timepoints $t_1 = 1,\ t_2 = 2,\ \ldots,\ t_k = t$

$$
\begin{aligned}
r_1(t) &= \mathbb{P}\left[T \leq t, \delta = 1\right] \\
&= \mathbb{P}\left[T = 1, \delta = 1\right] + \mathbb{P}\left[1 < T \leq 2, \delta = 1\right] + \ldots \\
&= \mathbb{P}\left[T = 1, \delta = 1\right] + \mathbb{P}\left[T = 2, \delta = 1 | T > 1\right] \mathbb{P}\left[T \geq 1\right] + \ldots \\
&= \lambda_{01}(1) + \lambda_{01}(2) S(1) + \ldots \\
&= \int_{s=0}^{t} \lambda_{01}(s) S(s-) ds
\end{aligned}
$$

where the all cause survival (no death nor infection) is:

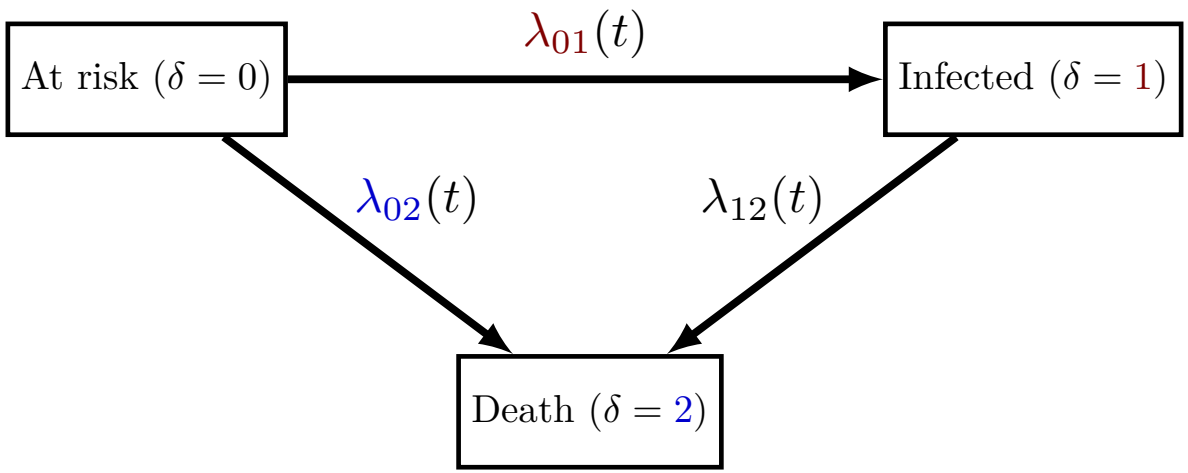
$$
\begin{aligned}
S(t) &= \left(1 - \lambda_{01}(1) - \lambda_{02}(1)\right)\left(1 - \lambda_{01}(2) - \lambda_{02}(2)\right)\ldots \\
&= \prod_{s=0}^{t} \left(1 - \lambda_{01}(s) ds - \lambda_{02}(s) ds\right)
\end{aligned}
$$

Re-defining the risk (2/2)

The "absolute" risk for the event of interest depends on the rate for the competing risks

$$r_1(t) = \lambda_{01}(1) + \lambda_{01}(2)\,(1 - \lambda_{01}(1) - \lambda_{02}(1)) + \ldots$$

$$= \int_{s=0}^{t} \lambda_{01}(s) \prod_{u=0}^{s-} (1 - \lambda_{01}(u)du - \lambda_{02}(u)du)\, ds$$



Classical mistakes (Andersen et al., 2012)

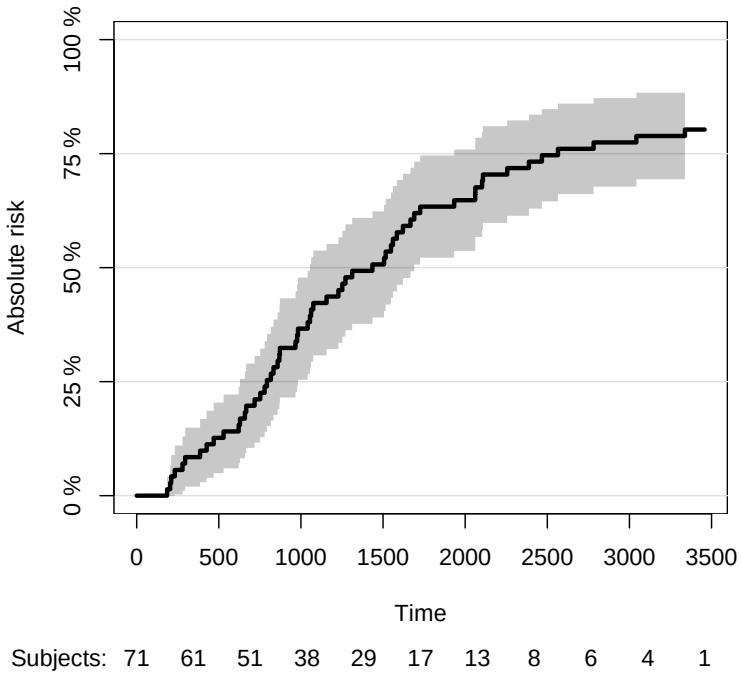
1. Treating competing events as censorings:
   - is conceptually wrong: risk had nobody died!
     $\rightarrow$ violate principle 3!
     $\rightarrow$ do not use Kaplan Meier!
   - gives wrong results: upwards biased estimate of the risk (since the event is no more prevented by death)

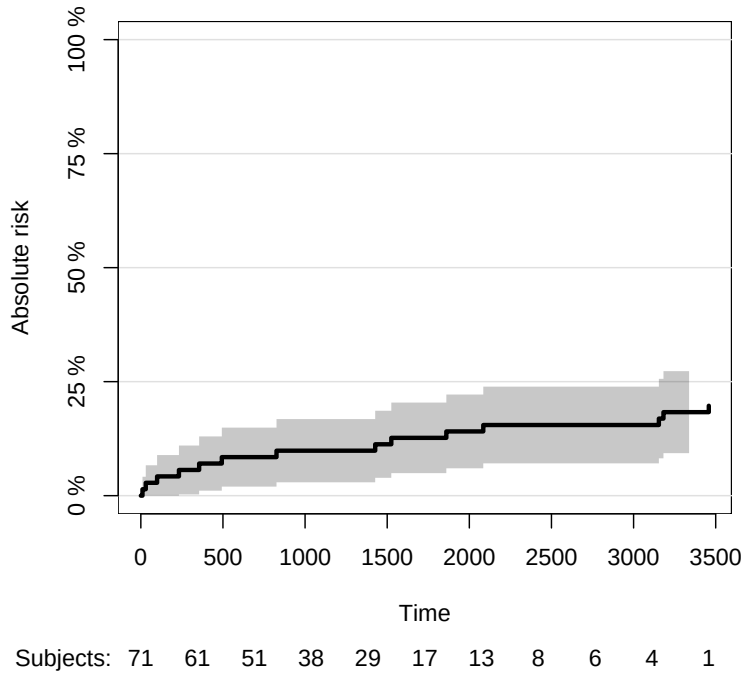
2. Only considering the event of interest:
   - incomplete picture: report the risk for each event

(by killing people a treatment may decrease the risk of stroke)

Aalen Johansen estimator

Risk estimator in presence of competing risk
and (independent) right-censoring






```
e.AJ <- prodlim(Hist(time, status) ~ 1, data = Melanoma2)
par(mfrow = c(1,2))
plot(e.AJ, cause = 1, title = "Cancer related death")
plot(e.AJ, cause = 2, title = "Death from other causes")
```

Exercise!

File *exercise-workshopEpi.R* (line 218-243)

Evaluate the 5-year risk of death for each cause with the Aalen Johansen estimator:

- in the manipulated dataset `Melanoma2`.
  Are the results surprising?
- in the original dataset `Melanoma`

Note: similar results can be obtained with the IPCW approach

What we have seen today

- ✔ 2 measures of disease frequency: risk & rate
  - time matters! From when until when?
  - risk - rate relationship (also with competing risks)
- ✔ Some classical mistakes
  - ever treated vs. never treated (immortal time bias)
  - exclude patients with censoring/competing risks
  - treat competing risks as censoring or no event
- ✔ 3 safety principles
  - Do not condition on the future
  - Do not condition on having reached an absorbing state
  - Stick to this world
- ✔ Handling treatment heterogeneity
  - complex model + G-formula
- ✔ Handling right-censoring & competing risks
  - modeling the rate and deducing the risk (KM,AJ)
  - re-weighting the observations (IPCW)

Take home message

Analyzing registry data is often challenging:

- partially observed outcome (censoring, competing risks)
- time varying exposure
- confounding, ...

A reasonnable approach goes as follow:

- **target**: precise description of the measure of disease frequency
- **ideal**: analysis had you had complete/balanced data
- **real**: what are the difficulties?
  what do we know or can assume:
  - about the censoring mechanism: IPCW
  - about the incidence rate: KM, AJ, Cox

Reference I

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology*, 41(3):861–870.

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.

Reference II

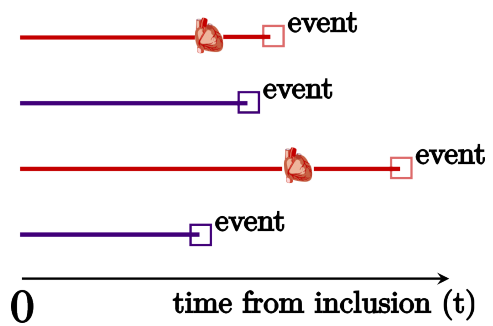
Jensen, H., Benn, C. S., Lisse, I. M., Rodrigues, A., Andersen, P. K., and Aaby, P. (2007). Survival bias in observational studies of the impact of routine immunizations on childhood survival. *Tropical Medicine & International Health*, 12(1):5–14.

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.

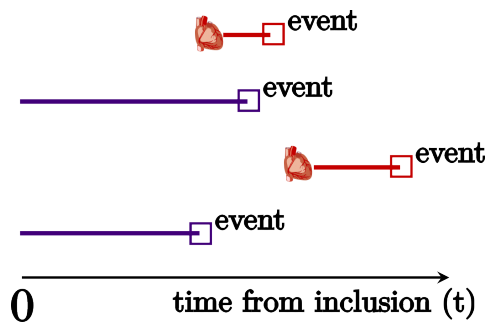
Satten, G. A. and Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210.

Immortal time bias (1/2)

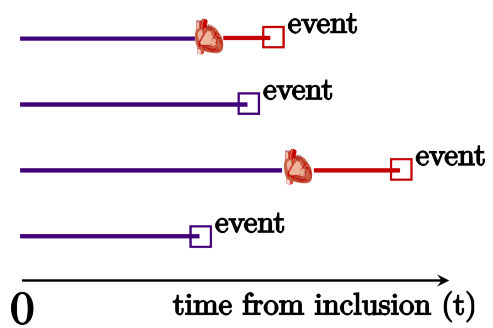
Solution 1:



Solution 2:



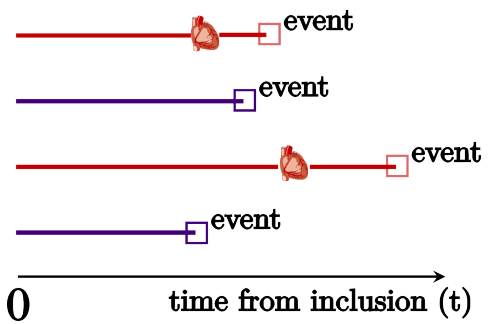
Solution 3:



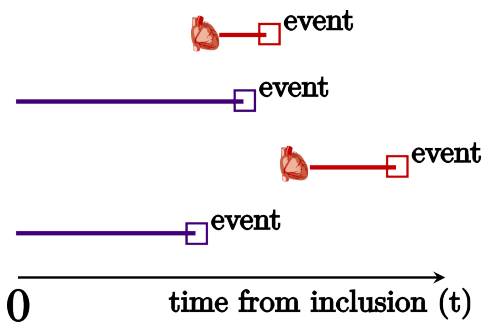
Immortal time bias (1/2)

✘ Solution 1:

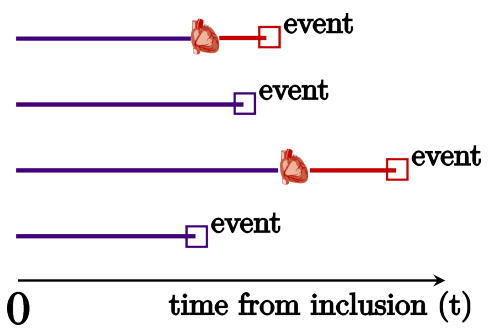
- irrealistic: use future information to define exposure
- immortal time bias: baseline-transplant



Solution 2:



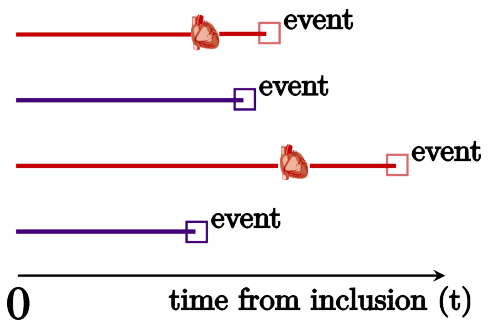
Solution 3:



Immortal time bias (1/2)

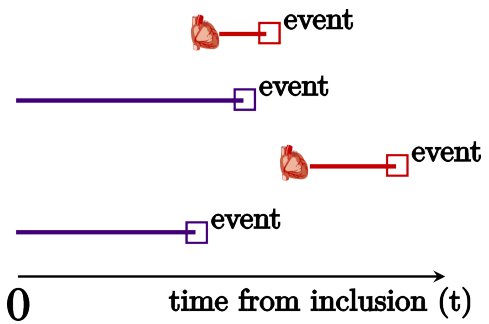
✘ Solution 1:

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

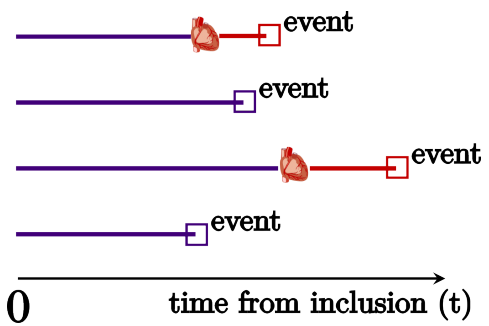


✘ Solution 2:

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



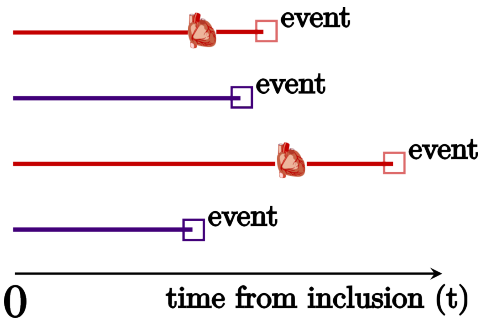
Solution 3:



Immortal time bias (1/2)

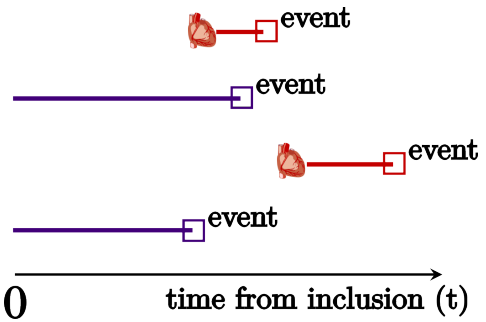
✘ Solution 1:

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



✘ Solution 2:

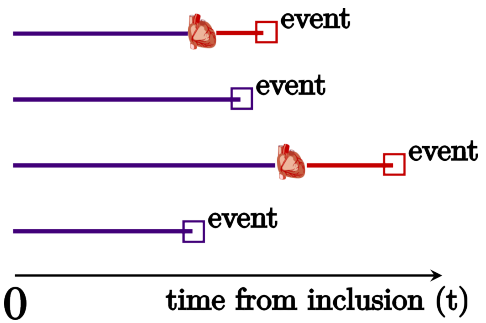
- irrealistic: use future information to remove data
- biased against no transplant



✔ Solution 3:

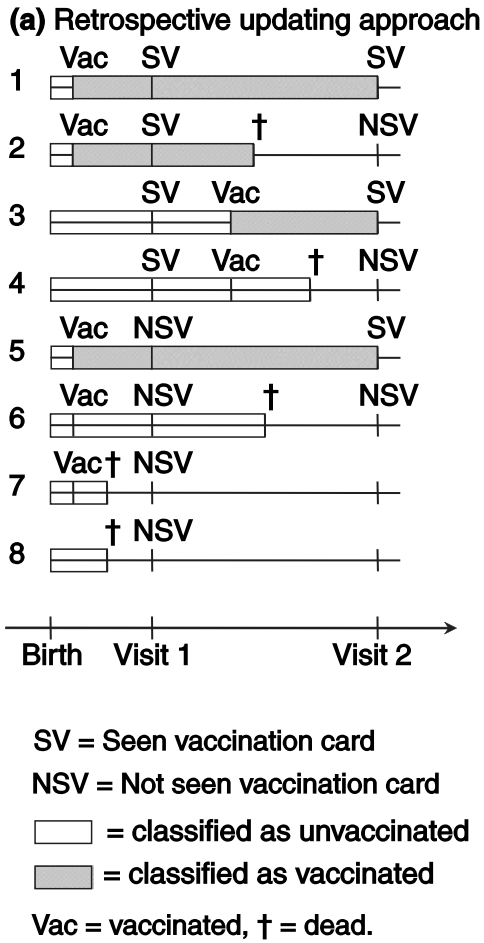
- realistic: time-varying exposure

😱 how to carry-out the analysis?



Immortal time bias (2/2)

From Jensen et al. (2007):

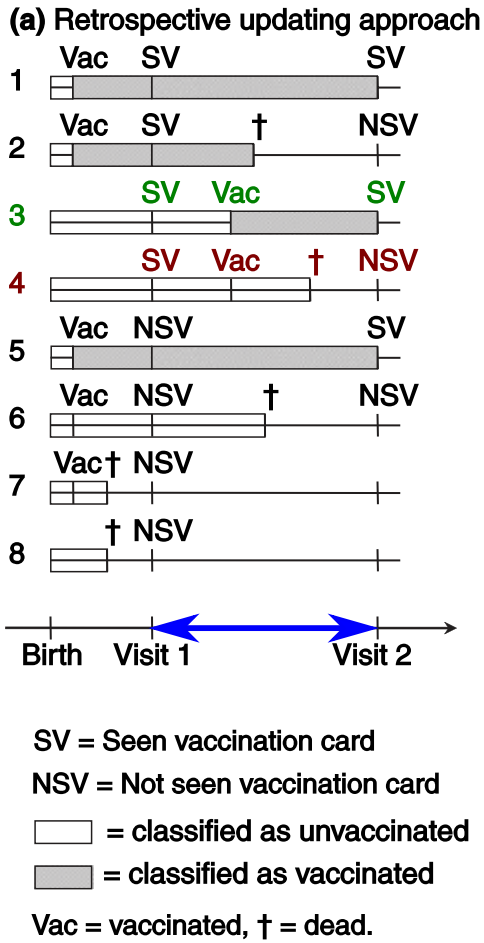


Retrospective updating approach

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.

Immortal time bias (2/2)

From Jensen et al. (2007):



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.