Lecture 13: Registry data analysis

Brice Ozenne[1,2] - brice.mh.ozenne@gmail.com

[1] Section of Biostatistics, Department of Public Health, University of Copenhagen

[2] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

16 March 2023

Recap'

- regression models for disease frequency
- regression models assessing exposure effect
- modeling time effects
- independence censoring assumption

Measures of disease frequency

- **Prevalence**: proportion of people with a disease

$$\hat{\pi} = \frac{\text{"number of people with the disease"}}{\text{"number of people"}}$$

- **Incidence rate**: frequency of disease occurrence over period τ
  ⚠ unit: time^{-1}, e.g. person-year.

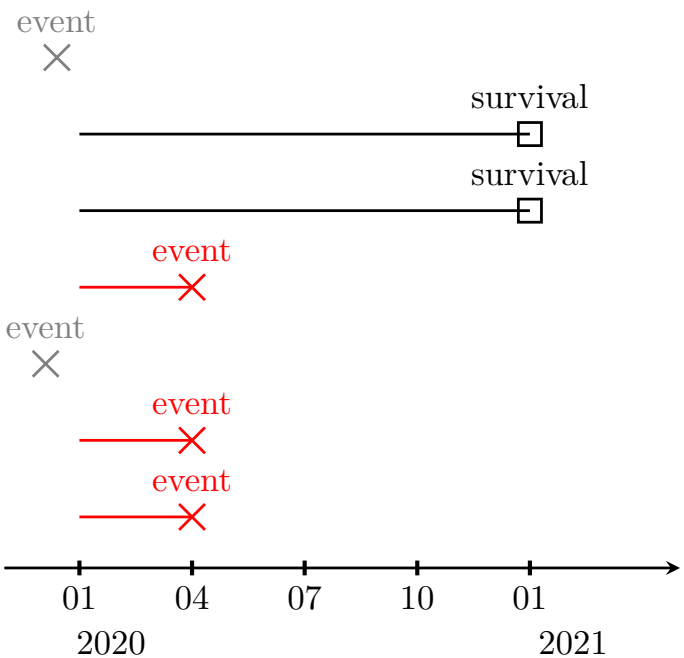
$$\widehat{\lambda} = \frac{\text{"number of new cases"}}{\text{"number of person-time at risk"}}$$

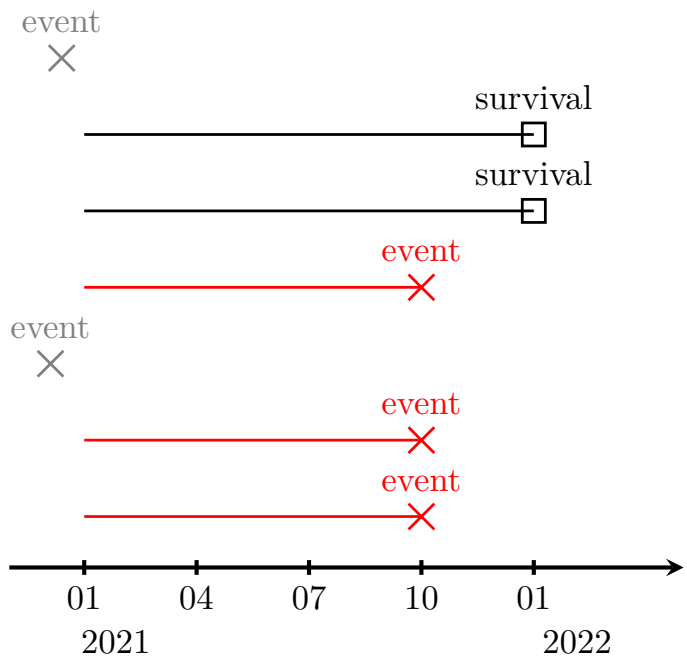
- **Risk**: probability of experiencing the disease before time τ

$$\widehat{r}_\tau = \frac{\text{"number of new cases"}}{\text{"number of person at risk"}}$$

Estimation "by hand"

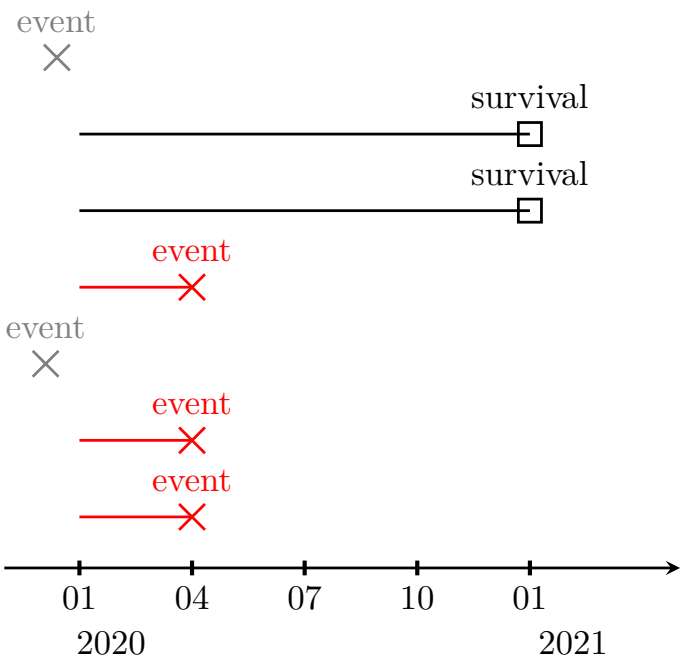
- **12-month risk**: $\hat{r}_{12}(2020) =$, $\hat{r}_{12}(2021) =$
- **6-month risk**: $\hat{r}_{6}(2020) =$, $\hat{r}_{6}(2021) =$
- **Incidence rate**: $\widehat{\lambda}(2020) =$ person-month,
  $\widehat{\lambda}(2021) =$ person-month

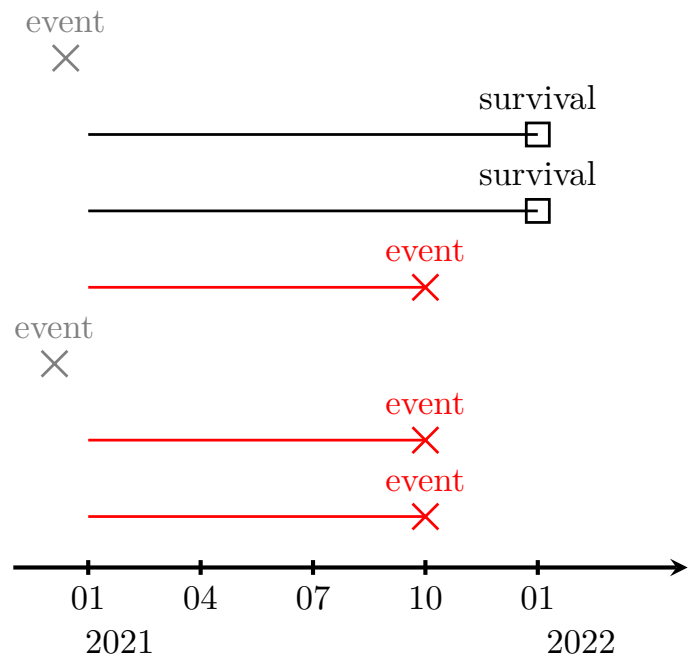




Estimation "by hand"

- **12-month risk**: $\hat{r}_{12}(2020) = 60\%$, $\hat{r}_{12}(2021) = 60\%$
- **6-month risk**: $\hat{r}_{6}(2020) = 60\%$, $\hat{r}_{6}(2021) = 0\%$
- **Incidence rate**: $\widehat{\lambda}(2020) = \frac{3}{3*3+2*12} \approx 0.0909$ person-month,
  $\widehat{\lambda}(2021) = \frac{3}{3*9+2*12} \approx .0588$ person-month





Regression models: basic logistic

```
df
```

```
id time status     id time status     id time status
 2   12      0      4    3      1      7    3      1
 3   12      0      6    3      1
```

Logistic model: $\log\left(\frac{r}{1-r}\right) = \alpha \iff r = \frac{1}{1+\exp(-\alpha)}$

```
e.prev <- glm(status ~ 1, data = df,
              family = binomial(link="logit"))

c(alpha_hat = as.double(coef(e.prev)),
  pi_hat = as.double(1/(1+exp(-coef(e.prev)))))
```

```
alpha_hat    pi_hat
0.4054651 0.6000000
```

Regression models: basic Poisson

Poisson model: $\log(\lambda) = \alpha \iff \lambda = \exp(\alpha)$

```
e.rate <- glm(status ~ 1, data = df,
              offset = log(time),
              family = poisson(link="log"))

c(alpha_hat = as.double(coef(e.rate)),
  lambda_hat = as.double(exp(coef(e.rate))))
```

```
  alpha_hat  lambda_hat
-2.39789527  0.09090909
```

Note: intuition for the offset $\widehat{\lambda} = \frac{\text{number of cases}}{\text{total time at risk}}$ denoted $\frac{D}{PY}$

$$\log(\lambda) = \alpha \iff \log(D) = 1 * \log(PY) + \alpha$$

Measures of association

We previously evaluated:

- **12-month risk**: $\hat{r}(2020) = 60\%$, $\hat{r}(2021) = 60\%$
- **Incidence rate**: $\widehat{\lambda}(2020) \approx 0.0909$, $\widehat{\lambda}(2021) \approx 0.0588$

  - **difference**: $\widehat{r}(2021) - \widehat{r}(2020) = 0$
    $\widehat{\lambda}(2021) - \widehat{\lambda}(2020) = -0.384$ person-month

  - **ratio**: $\frac{r(2021)}{r(2020)} = 1$
    $\frac{\lambda(2021)}{\lambda(2020)} = 0.647$

  - **odd ratio**: $\left(\frac{r(2021)}{1-r(2021)}\right) \Big/ \left(\frac{r(2020)}{1-r(2020)}\right) = 1.5/1.5 = 1$

Parametrisation of the logistic model

Logistic model: $\log\left(\frac{r(\text{year})}{1-r(\text{year})}\right) = \alpha + \beta * \text{ year}$

- in 2020: $\log\left(\frac{r(2020)}{1-r(2020)}\right) = \log\left(\Omega(2020)\right) = \alpha$
- in 2021: $\log\left(\frac{r(2021)}{1-r(2021)}\right) = \log\left(\Omega(2021)\right) = \alpha + \beta$

So $\Omega(2020) = \exp(\alpha)$
$\Omega(2021) = \exp(\alpha + \beta)$
and $OR = \frac{\Omega(2021)}{\Omega(2020)} = \exp(\beta)$

Parametrisation of the logistic model

Logistic model: $\log\left(\frac{r(\text{year})}{1-r(\text{year})}\right) = \alpha + \beta * \text{ year}$

- in 2020: $\log\left(\frac{r(2020)}{1-r(2020)}\right) = \log\left(\Omega(2020)\right) = \alpha$
- in 2021: $\log\left(\frac{r(2021)}{1-r(2021)}\right) = \log\left(\Omega(2021)\right) = \alpha + \beta$

So $\Omega(2020) = \exp(\alpha)$
$\Omega(2021) = \exp(\alpha + \beta)$
and $OR = \frac{\Omega(2021)}{\Omega(2020)} = \exp(\beta)$

⚠ not feasible in presence of right-censoring:

- Cox/Poisson regression
- IPCW logistic

Measures of association - logistic model

```
df2
```

```
id year time status     id year time status     id year time status
 1 2020   12      0      4 2020    3      1      3 2021    9      1
 2 2020   12      0      5 2020    3      1      4 2021    9      1
 3 2020    3      1      1 2021   12      0      5 2021    9      1
                         2 2021   12      0
```

Output from the logistic model:

```
e.OR <- glm(status ~ year, data = df2,
            family = binomial(link = "logit"))
exp(coef(e.OR))
```

```
(Intercept)    year2021
        1.5         1.0
```

Parametrisation of the Poisson model

Poisson model: $\log(\lambda(year)) = \alpha + \beta * \text{year}$

- in 2020: $\log(\lambda(2020)) = \alpha$
- in 2021: $\log(\lambda(2021)) = \alpha + \beta$

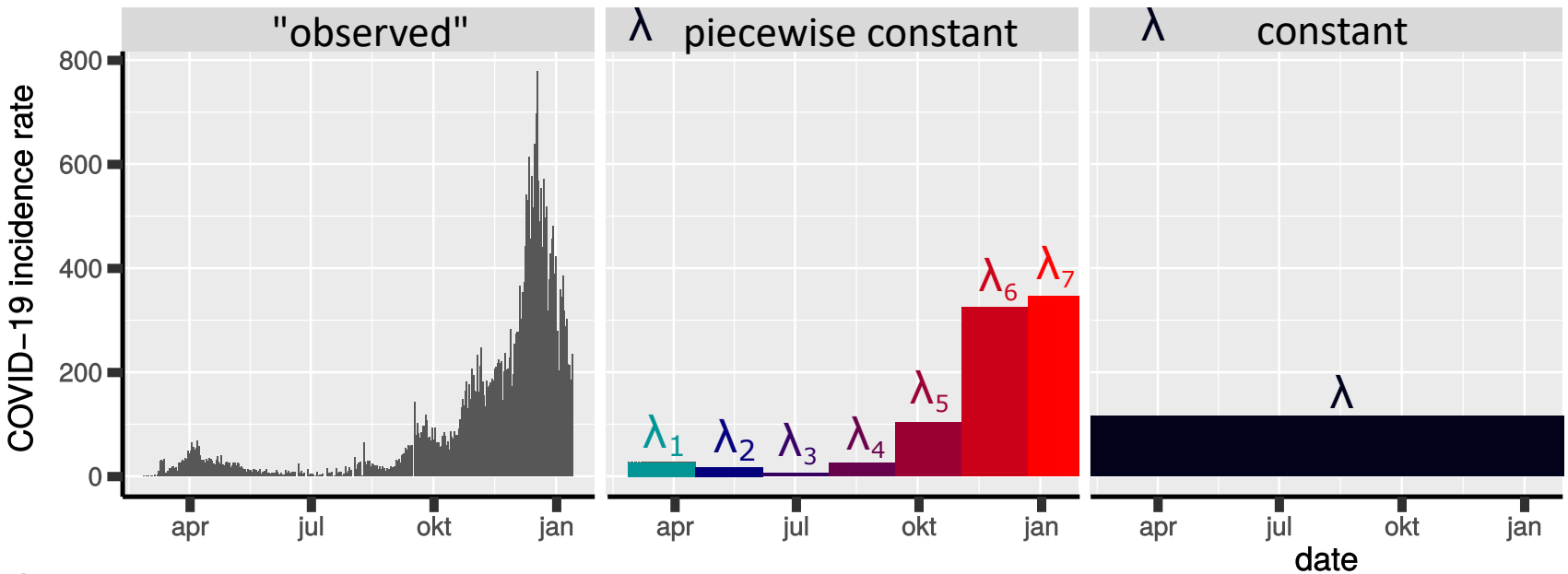
So $\frac{\lambda(2021)}{\lambda(2020)} = \exp(\beta)$

```
e.RR <- glm(status ~ year, data = df2,
            offset = log(time),
            family = poisson(link = "log"))
exp(coef(e.RR))
```

```
(Intercept)    year2021
 0.09090909  0.64705882
```

Handling time varying hazard

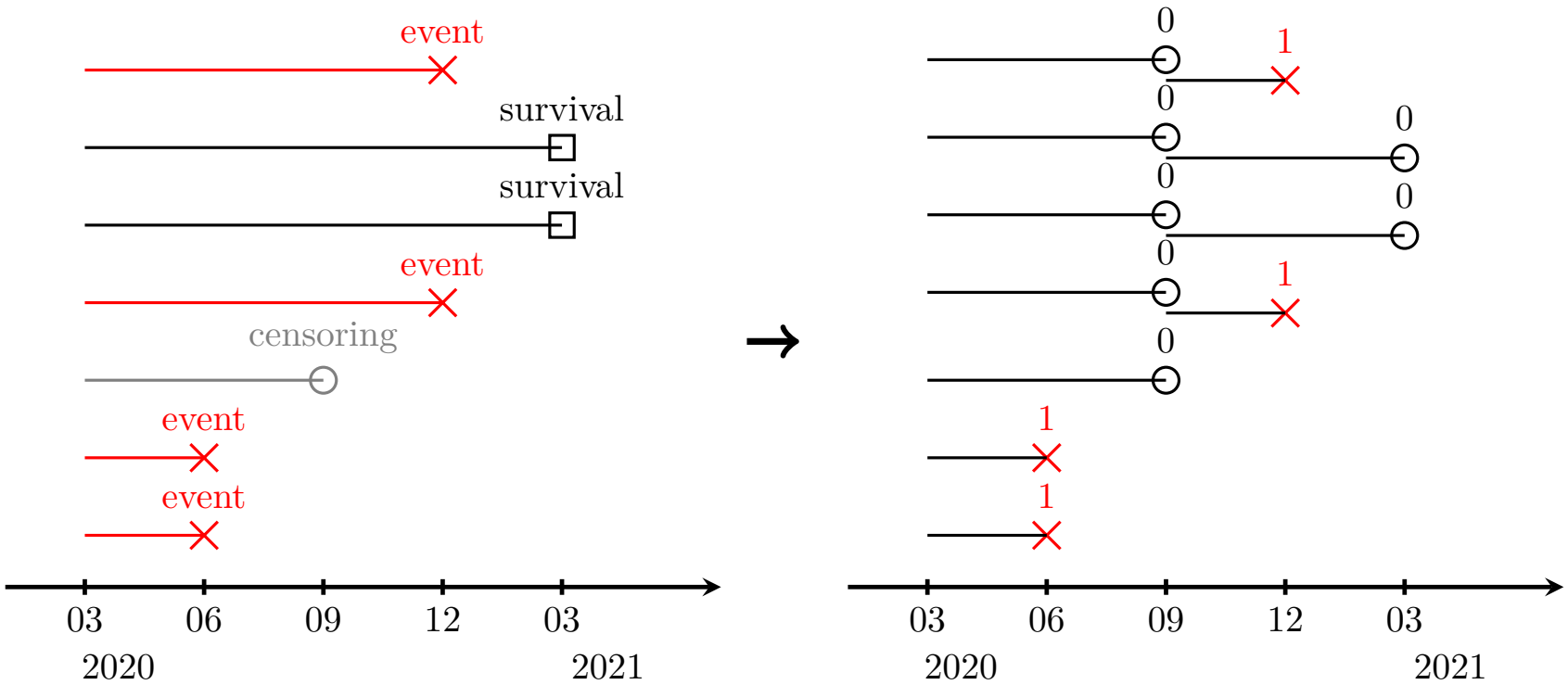
The "simple" Poisson model is often unrealistic:



Solutions:

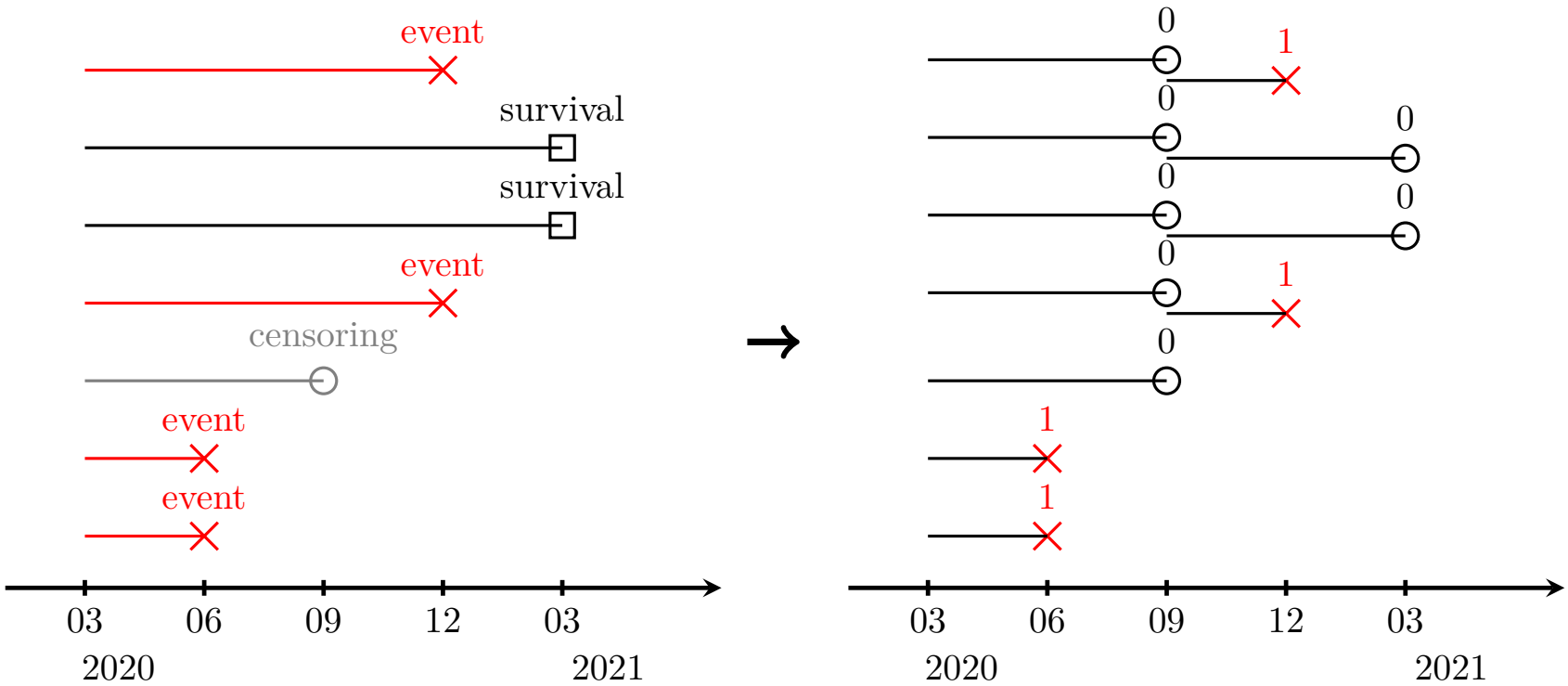
- time-splitting + Poisson: assumes piecewise constant hazard
- Cox model: no assumption on the shape of λ (semi-parametric estimator)

Time varying hazard - by hand



- $\lambda_{\text{per year}}(03/2020 - 09/2020) =$
- $\lambda_{\text{per year}}(09/2020 - 03/2021) =$

Time varying hazard - by hand



- $\lambda_{\text{per year}}(03/2020 - 09/2020) = \frac{2}{2*3+5*6} \approx 0.056$
- $\lambda_{\text{per year}}(09/2020 - 03/2021) = \frac{2}{2*3+2*6} \approx 0.111 \approx 2 * 0.056$

Time varying hazard - via a Poisson model

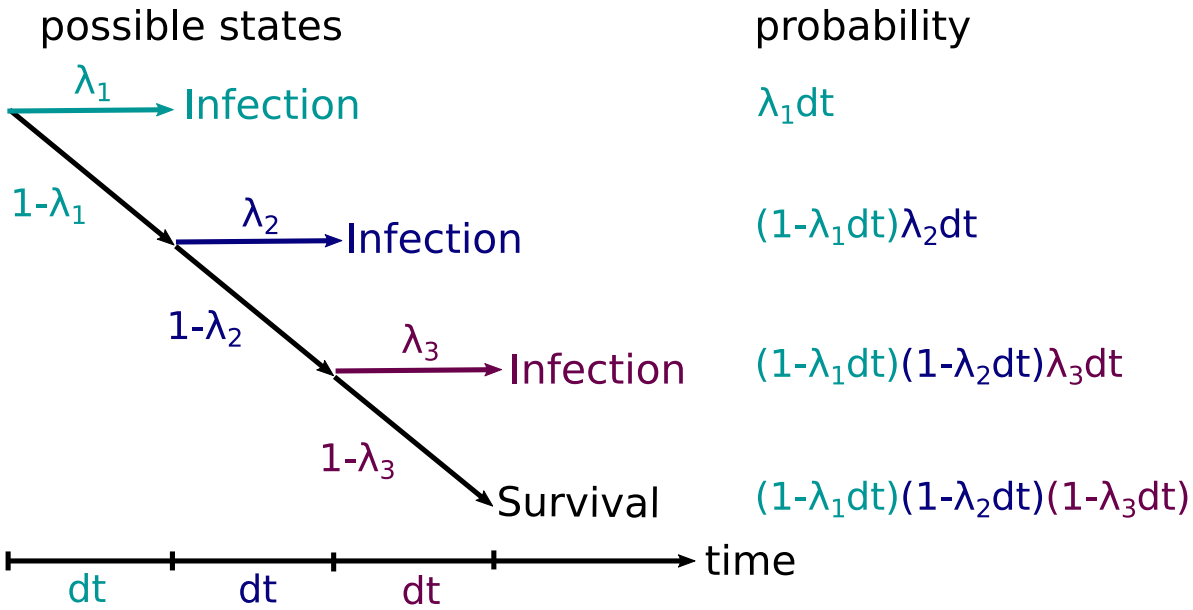
```
df3
```

id	time	status	period	id	time	status	period
1	6	0	1	7	3	1	1
2	6	0	1	1	3	1	2
3	6	0	1	2	6	0	2
4	6	0	1	3	6	0	2
5	6	0	1	4	3	1	2
6	3	1	1				

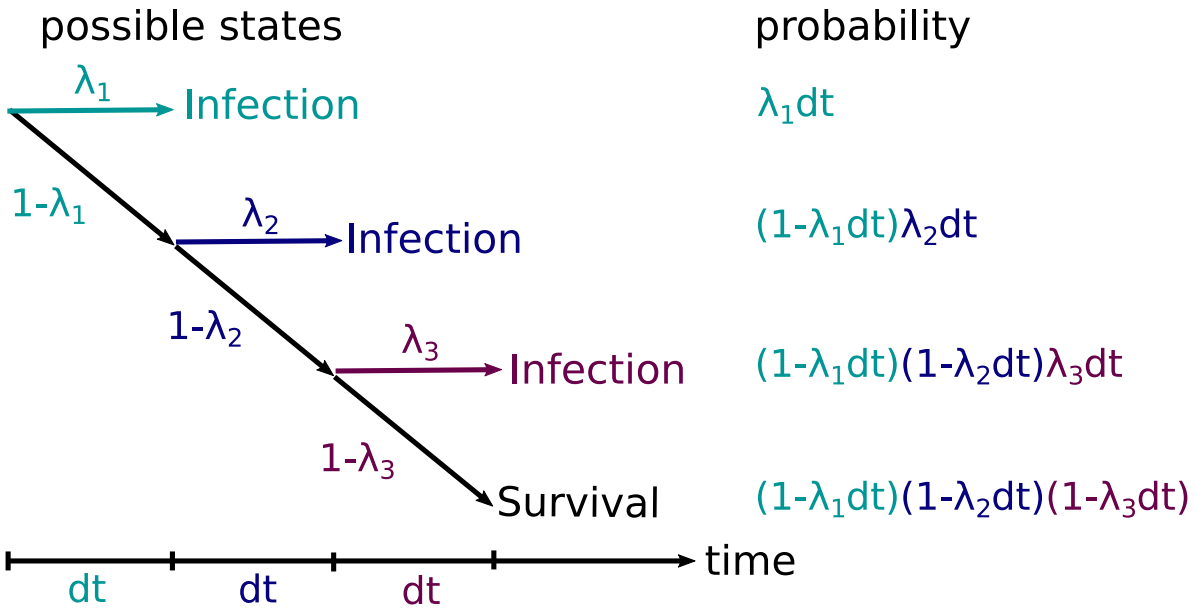
```
e.rateV <- glm(status ~ period, data = df3,
               offset = log(time),
               family = poisson(link="log"))
exp(coef(e.rateV))
```

```
(Intercept)     period2
 0.05555556  2.00000000
```

From the hazard to the survival



From the hazard to the survival

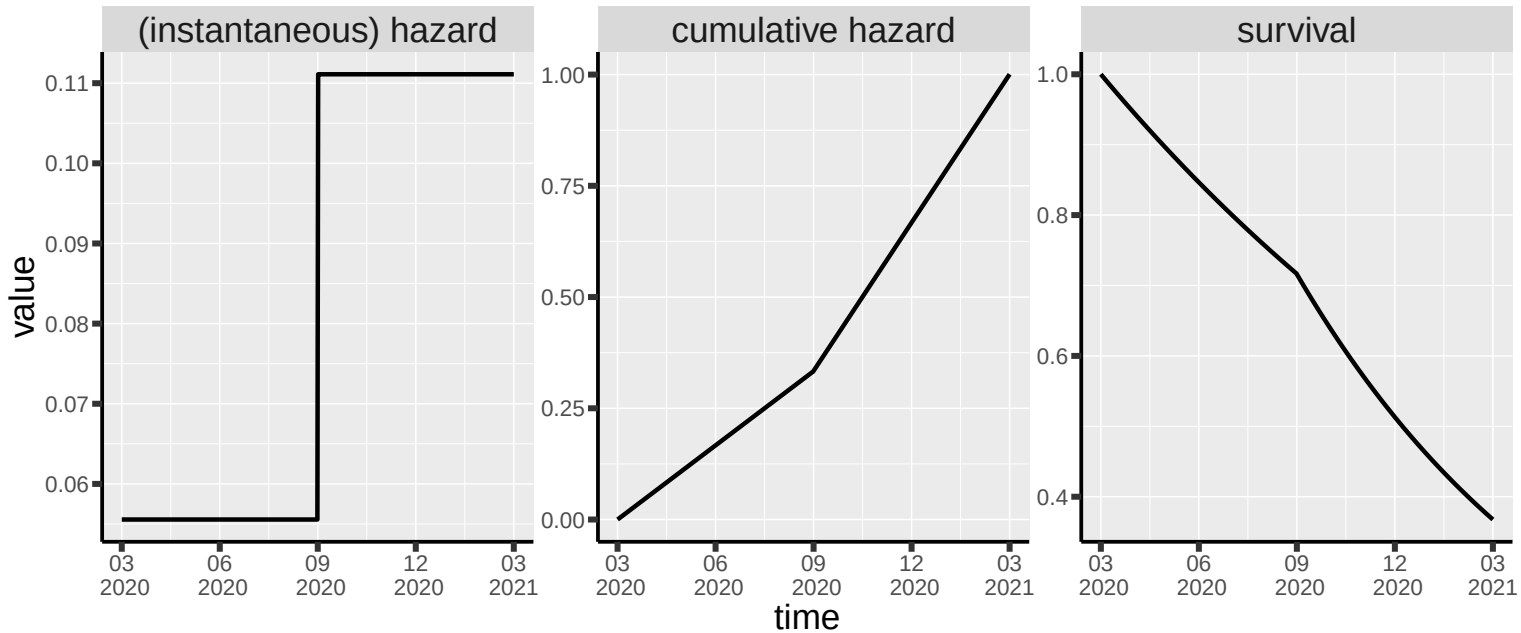


The 1-year risk of infection is:

$$r_{1\text{ year}} = 1 - S(t) = 1 - (1 - \lambda_1 dt)(1 - \lambda_2 dt)\ldots(1 - \lambda_7 dt)$$
$$\approx 1 - \exp(-(\lambda_1 + \lambda_2 + \ldots + \lambda_7)dt)$$

where $S(t)$ is the survival (i.e. staying infection free).
Approximation only accurate for small time intervals ($\lambda . dt \ll 1$).

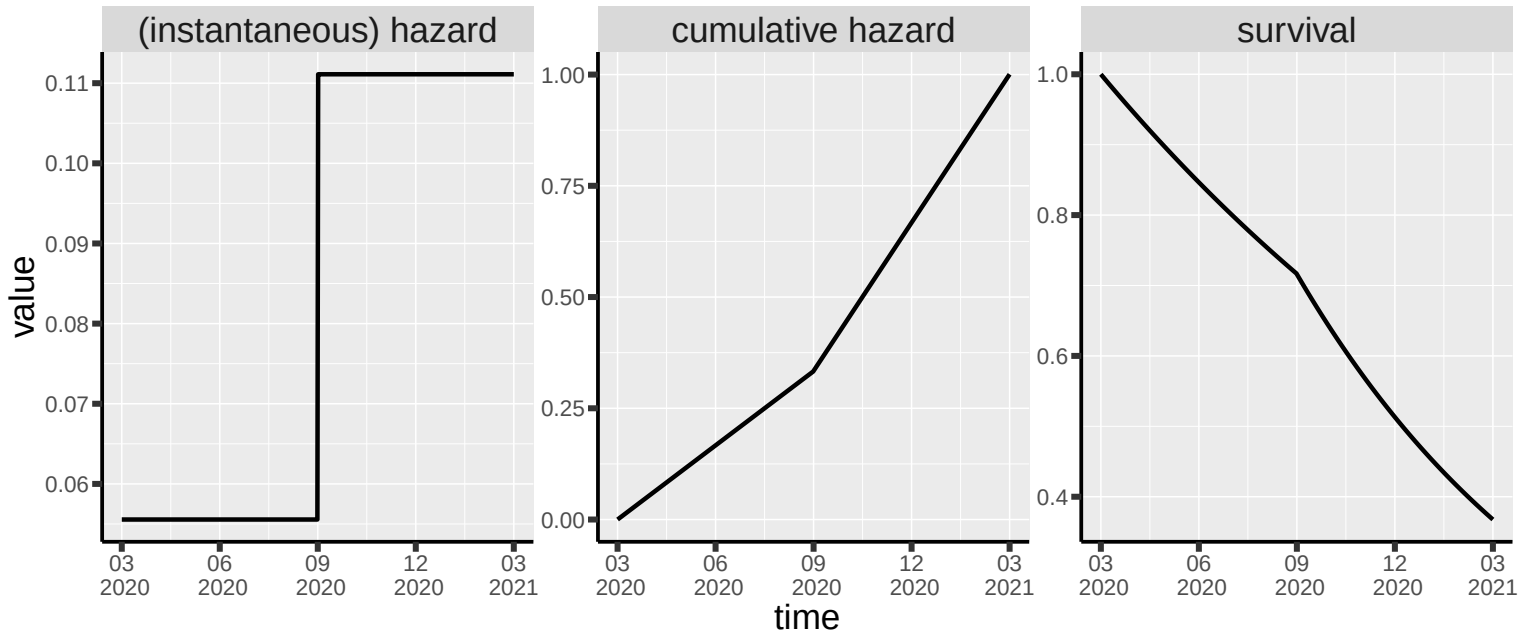
Hazard, cumulative hazard, and survival



Before 09/2020 (i.e. time ≤ 6);

- $\lambda(\text{time}) \approx 0.056$
- $\Lambda(\text{time}) = \int_{s=0}^{\text{time}} \lambda_{\text{per year}}(s)ds \approx 0.056 \times \text{time}$
- $S(\text{time}) \approx \exp\left(-\Lambda(\text{time})\right) \approx \exp(-0.056 \times \text{time})$

Hazard, cumulative hazard, and survival



After 09/2020 (i.e. time > 6);

- $\lambda(\text{time}) \approx 0.111$
- $\Lambda(\text{time}) \approx 0.056 \times 6 + 0.111 \times (\text{time} - 6)$
- $S(\text{time}) \approx \exp(-0.056 \times 6 - 0.111 \times (\text{time} - 6))$

Why using Poisson/Cox regression?

Difficult to extend "by hand" calculations to deal with:

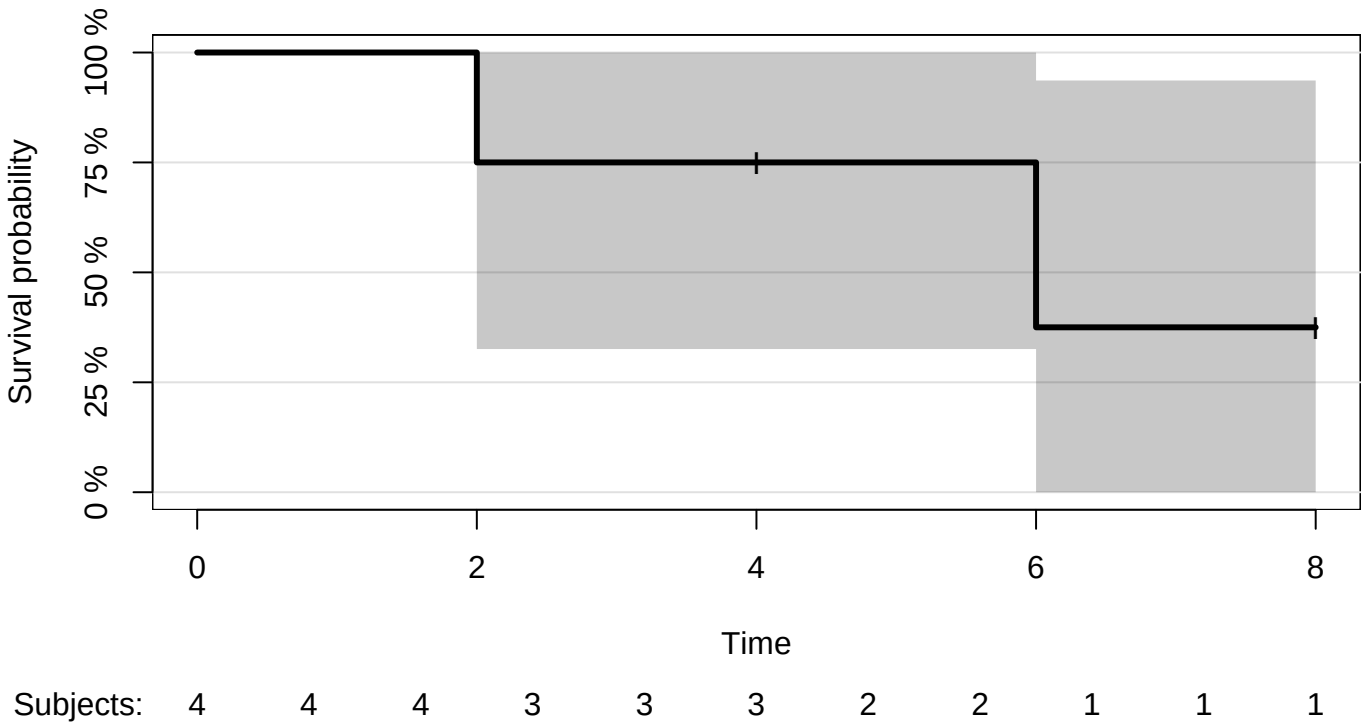
- censoring
- confounding
- time varying hazards (i.e. time varying incidence rates)

$\rightarrow$ model the incidence λ to obtain the risk r

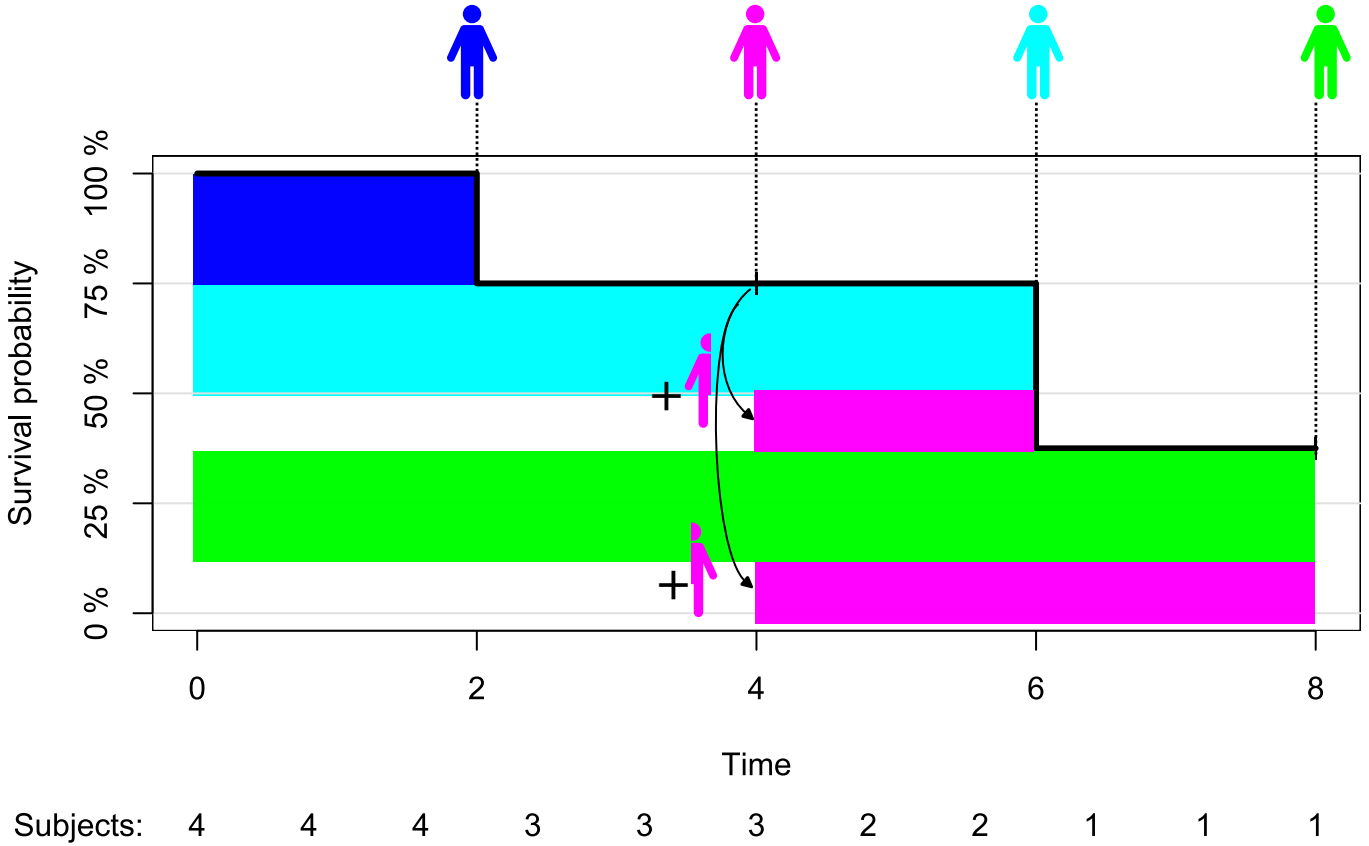
Cox vs. Poisson:

- Cox is a convenient and good "default" model.
- Poisson is useful when exposure/covariate effects are time varying.

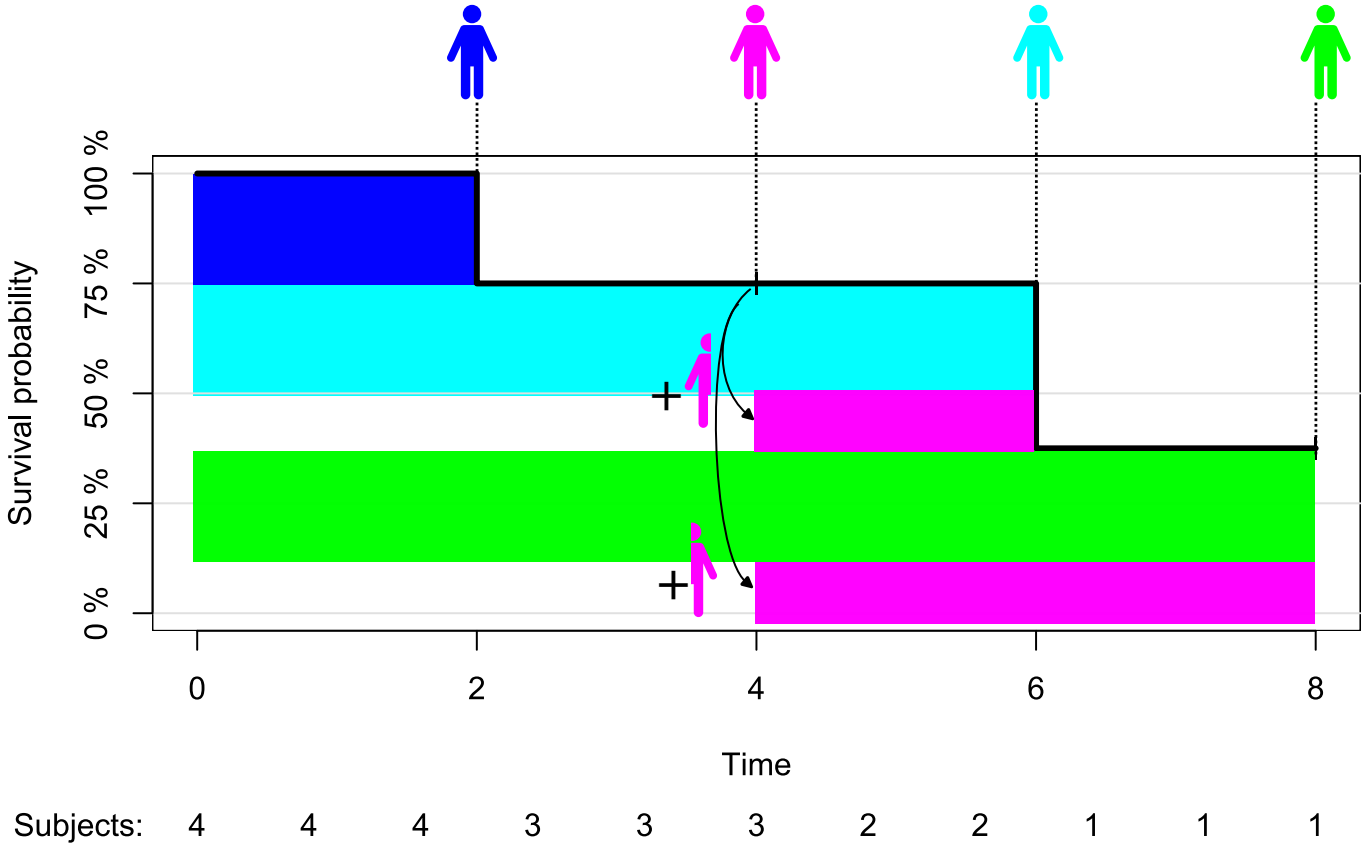
Another view at Kaplan Meier



Another view at Kaplan Meier



Another view at Kaplan Meier



- 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

In presence of right-censoring, we often assume that:

- survival times (T) and censorship times (C) are independent
- conditional on the covariates (X)

Said otherwise:

- within age and vaccine subgroups, subjects who are not censored at time t should be representative of all the subjects who remained at risk.

How critical is that assumption?

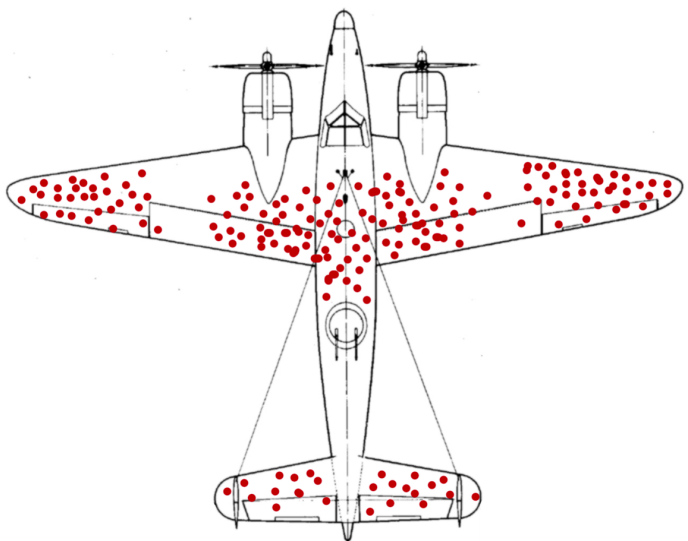
How to armor the planes?

During WW2 (1943), the army asked a research group where and how much armor to put on the plane:

- it protects planes from the bullets of enemy fighters
- but makes the plane heavier, less maneuverable

Among 400 planes,
380 have returned:

- 320 with no hit
- 32 with 1 hit
- 20 with 2 hit
- 8 with 3 or more hit



Source https://en.wikipedia.org/wiki/Survivorship_bias#/media/File:Survivorship-bias.png

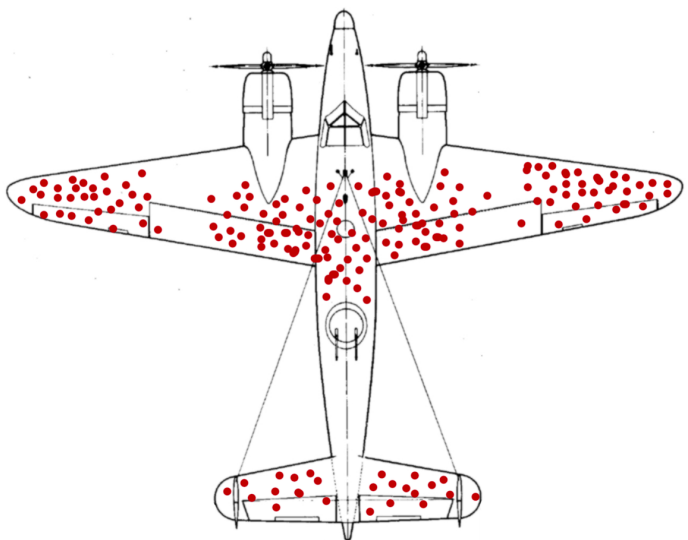
How to armor the planes?

During WW2 (1943), the army asked a research group where and how much armor to put on the plane:

- it protects planes from the bullets of enemy fighters
- but makes the plane heavier, less maneuverable

Among 400 planes,
380 have returned:

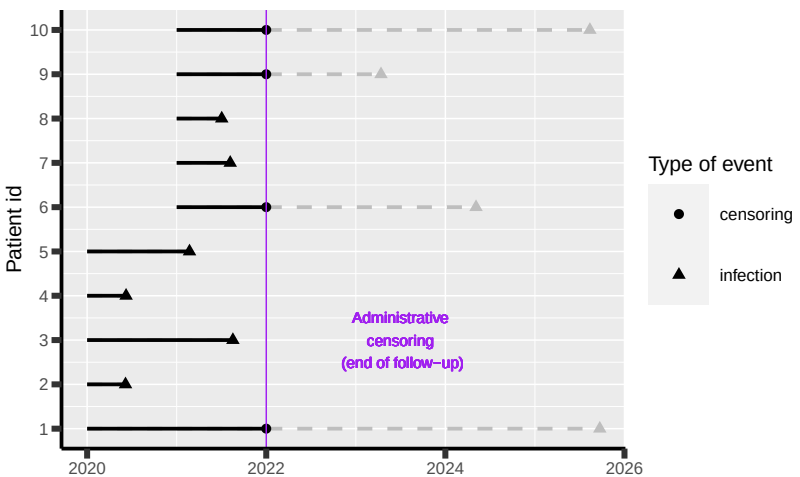
- 320 with no hit
- 32 with 1 hit
- 20 with 2 hit
- 8 with 3 or more hit

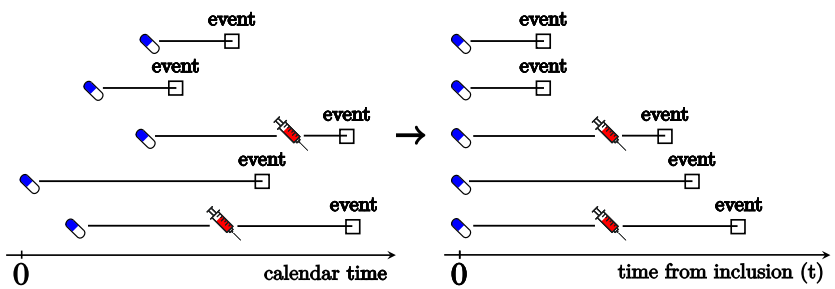


Source https://en.wikipedia.org/wiki/Survivorship_bias#/media/File:Survivorship-bias.png

"The armor, doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines." (Abraham Wald)

Registry data





Registry data

In Denmark, data about date of medicine purchase, hospital admission, or diagnostic of certain diseases can be found in the danish national registry.

- cover the danish population (leaving in Denmark) and foreigners living in Denmark.
- different registries for different types of information (prescription, psychiatry, . . .) that started at different dates.

What specificities of registry data can you see 

- implication for the statistical analysis

Registry data - some specificities

- typically **observational**: many covariates to be adjusted for
  `e.g. avoid confounding by indication`
- ⚠ follow-up time is subject dependent
  `e.g. young people have short follow-up time`
- ⚠ date of inclusion in the registry may not be medically relevant
  `e.g. date of emigration to Denmark`

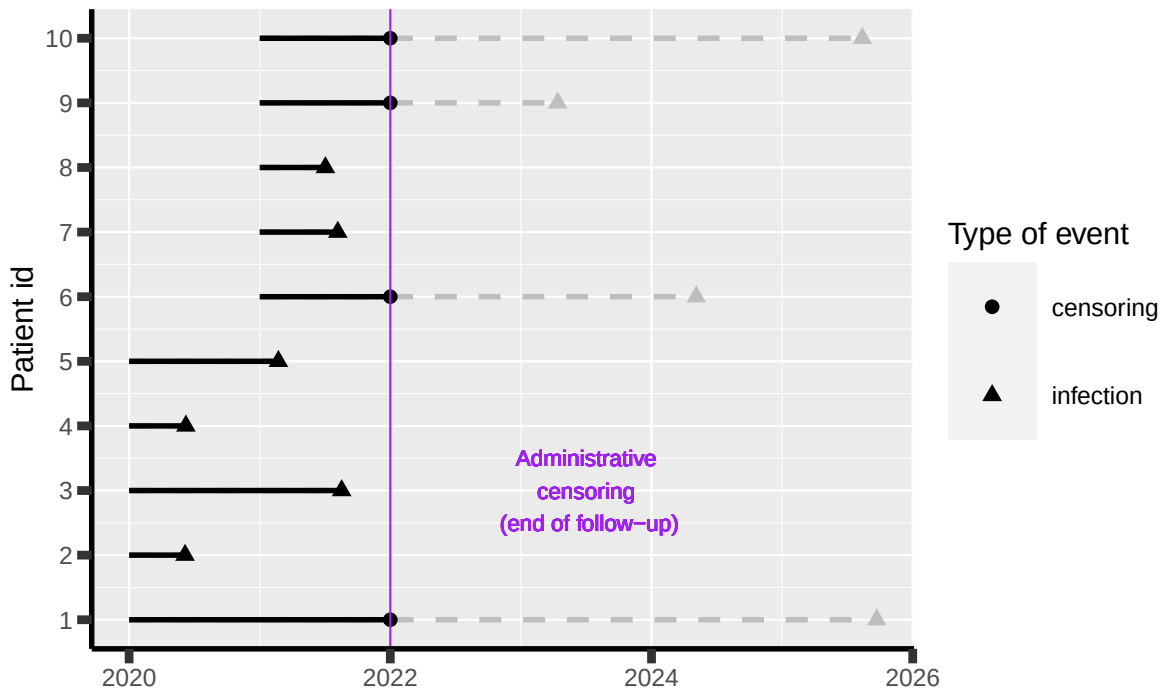
- **large dataset**: CIs typically more informative than p-values

- **long follow-up time**: outcome may not be observable due to other events `e.g. death as competing risk`

- **time varying exposure**: switch of treatment for unknown reasons `e.g. previous treatment was not working, or no more available, or the switch was planned`

What is wrong with this analysis?

Risk of death between start and end of follow-up: 60%

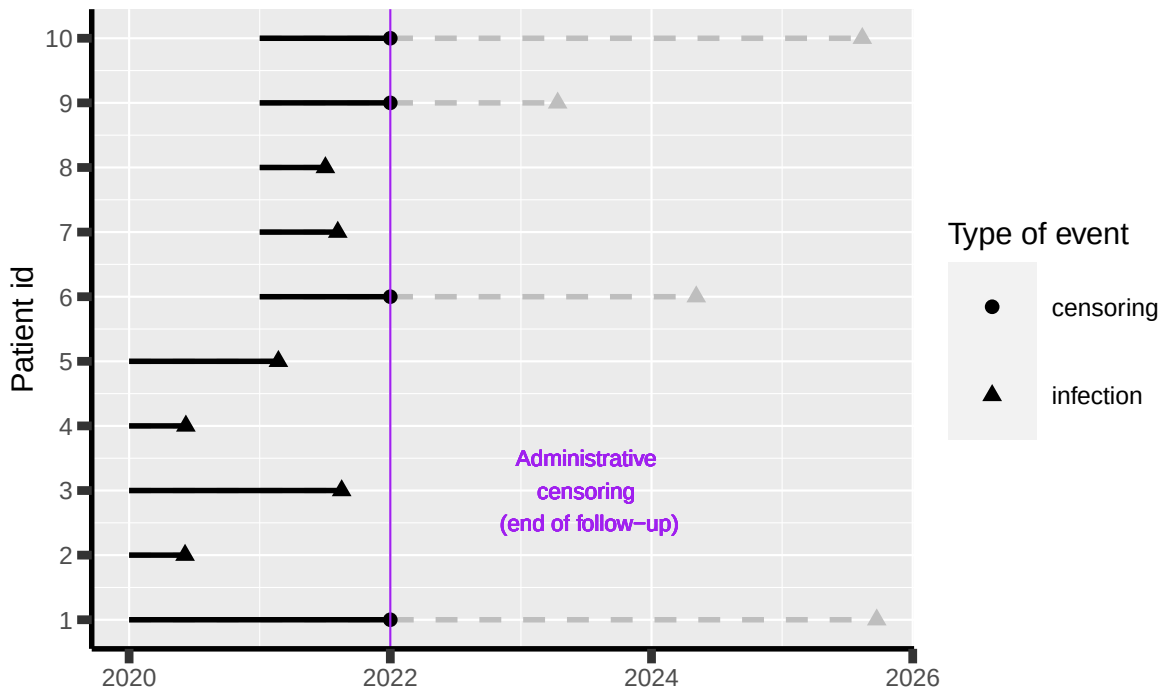


What is wrong with this analysis?

Risk of death between start and end of follow-up: 60%

⚠ no clear interpretation! Mix of 1 year risk (40%) and 2 year risk (80%)

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

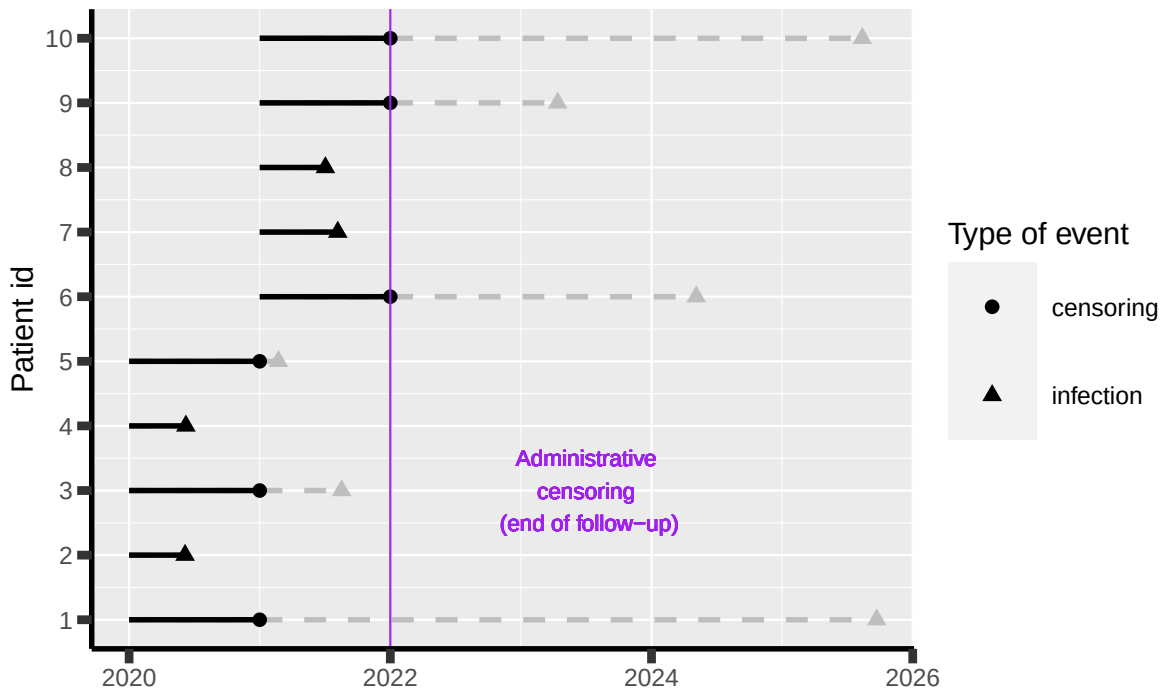


What is wrong with this analysis?

Risk of death between start and end of follow-up: 60%

⚠ no clear interpretation! Mix of 1 year risk (40%) and 2 year risk (80%)

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



Standardization

- motivation and intuition

- examples

Back to the BCG study

	age	[0,10]	[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

A different risk difference for each age group[1]:

- $\theta_1 = -5.929\%$ $n_1 = 269$
- $\theta_2 = -1.861\%$ $n_2 = 3181$
- $\theta_3 = 0.63\%$ $n_2 = 1810$

[1] age groups are not realistic - just illustrate age-dependent vaccine effects

What to do?

- ignore the interaction (easy to report but probably wrong)
- keep the interaction (difficult to report, less likely wrong)

What to do?

- ignore the interaction (easy to report but probably wrong)
- keep the interaction (difficult to report, less likely wrong)
- keep the interaction and compute an 'average' effect:

What to do?

- ignore the interaction (easy to report but probably wrong)
- keep the interaction (difficult to report, less likely wrong)
- keep the interaction and compute an 'average' effect:

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

Here for the risk difference:

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

Exercise (bissau study)

age	No vaccine	Vaccine	Number of individuals
0	$r_{0,\text{no}} = 4.29\%$	$r_{0,\text{yes}} = 2.21\%$	$n_{0,\text{no}} = 637,\ n_{0,\text{yes}} = 237$
1	$r_{1,\text{no}} = 5.02\%$	$r_{1,\text{yes}} = 2.77\%$	$n_{1,\text{no}} = 421,\ n_{1,\text{yes}} = 468$
2	$r_{2,\text{no}} = 3.82\%$	$r_{2,\text{yes}} = 1.87\%$	$n_{2,\text{no}} = 321,\ n_{2,\text{yes}} = 598$
ATE	$r_{.,\text{no}} =$	$r_{.,\text{yes}} =$	$n_{.,\text{no}} = 1379,\ n_{.,\text{yes}} = 1303$

$$r_{.,\text{no}} =$$

$$r_{.,\text{yes}} =$$

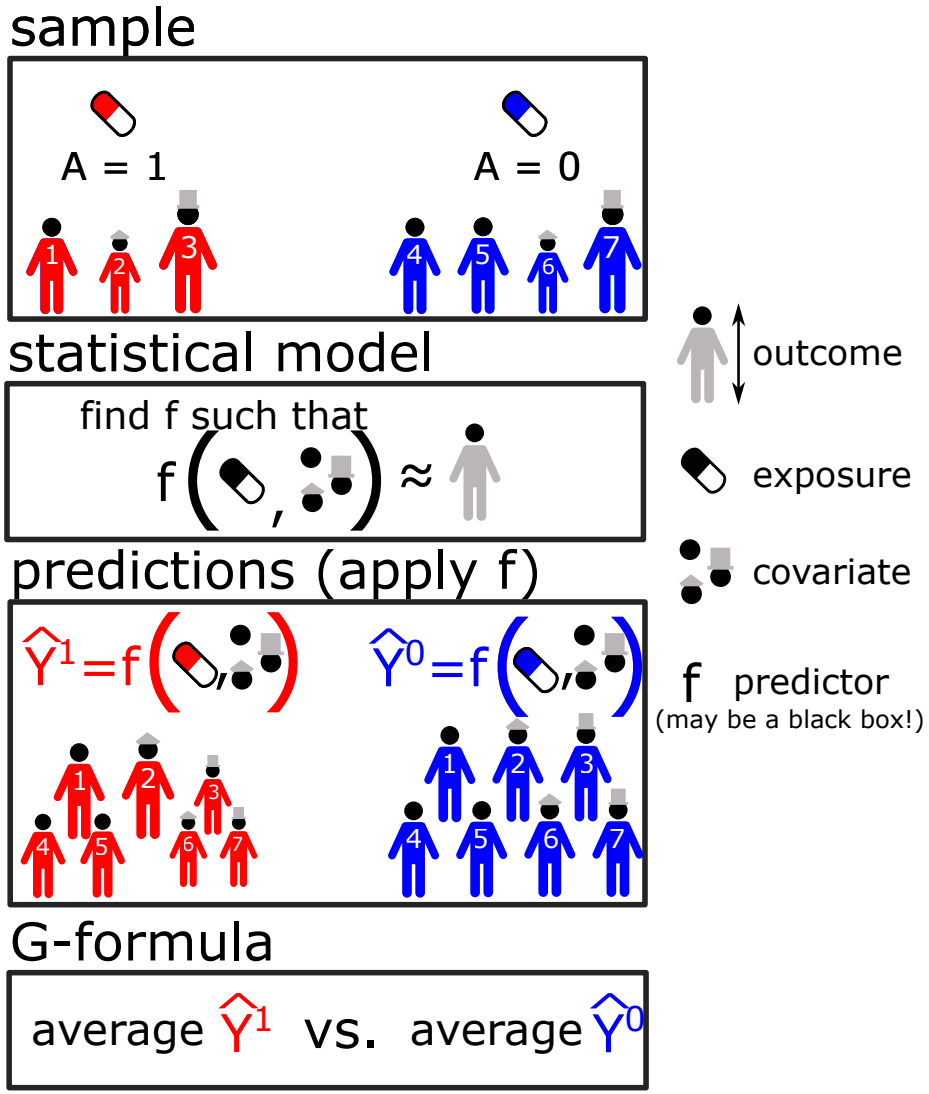
$$\Psi = r_{.,\text{yes}} - r_{.,\text{no}}$$

Exercise (bissau study)

age	No vaccine	Vaccine	Number of individuals
0	$r_{0,\text{no}} = 4.29\%$	$r_{0,\text{yes}} = 2.21\%$	$n_{0,\text{no}} = 637,\ n_{0,\text{yes}} = 237$
1	$r_{1,\text{no}} = 5.02\%$	$r_{1,\text{yes}} = 2.77\%$	$n_{1,\text{no}} = 421,\ n_{1,\text{yes}} = 468$
2	$r_{2,\text{no}} = 3.82\%$	$r_{2,\text{yes}} = 1.87\%$	$n_{2,\text{no}} = 321,\ n_{2,\text{yes}} = 598$
ATE	$r_{.,\text{no}} = 4.37\%$	$r_{.,\text{yes}} = 2.28\%$	$n_{.,\text{no}} = 1379,\ n_{.,\text{yes}} = 1303$

$$
\begin{aligned}
(p_1, p_2, p_3) &= \left(\frac{637+237}{1379+1303}, \frac{421+468}{1379+1303}, \frac{321+598}{1379+1303}\right) \\
&= (32.59\%, 33.15\%, 34.27\%) \\
r_{.,\text{no}} &= 32.59\% * 4.29\% + 33.15\% * 5.02\% + 34.27\% * 3.82\% \\
r_{.,\text{yes}} &= 32.59\% * 2.21\% + 33.15\% * 2.77\% + 34.27\% * 1.87\% \\
\Psi &= r_{.,\text{yes}} - r_{.,\text{no}} \approx 2.09\%
\end{aligned}
$$

Extension to continuous covariates



Comments

The average treatment effect (ATE) **depends on the population**. This is not the case with age-specific vaccine effects nor with regression coefficients.

Positivity assumption: any patient has a non-0 possibility to received any treatment.

In a linear regression $Y = \alpha + \beta E + \gamma age + \delta E \times age$,
the ATE is a weighted average of the covariate-specific treatment effect (β, δ).
In a non-linear model $logit(p) = \alpha + \beta E + \gamma age + \delta E \times age$,
the ATE may depend on all coefficients!

Example: bissau study

```
## get data
bissau <- read.table("https://bozenne.github.io/doc/
    Teaching/bissau.txt", header=TRUE)
bissau$status <- bissau$fupstatus=="dead"
bissau$agem <- as.factor(bissau$agem)
## Fit a statistical model (survival):
library(survival)
e.cox <- coxph(Surv(fuptime, status) ~ agem / bcg,
               data = bissau, x = TRUE)
exp(coef(e.cox))
```

```
      agem1       agem2       agem3       agem4       agem5
  1.1745375   0.8876400   1.1396365   0.8129175   0.4364516
agem0:bcgyes agem1:bcgyes agem2:bcgyes agem3:bcgyes agem4:bcgyes
  0.5088731   0.5462568   0.4850764   0.7142407   1.3237650
agem6:bcgyes
  0.1657378
```

Example: bissau study

```
## Predict the counterfactual risks:
library(riskRegression)
grid0 <- data.frame(bcg = "no", agem = factor(0:2))
grid1 <- data.frame(bcg = "yes", agem = factor(0:2))
r0 <- predictRisk(e.cox, newdata = grid0, time = 150)
r1 <- predictRisk(e.cox, newdata = grid1, time = 150)
round(100*data.frame(no = r0, yes = r1),2)
```

```
    no  yes
1 4.29 2.21
2 5.02 2.77
3 3.82 1.87
```

Standardization "by hand"

```
## Predict the counterfactual risks (riskRegression):
bissau0 <- bissau[bissau$agem %in% 0:2,]
bissau0$bcg <- "no"
r0 <- predictRisk(e.cox, newdata = bissau0, time = 150)

bissau1 <- bissau[bissau$agem %in% 0:2,]
bissau1$bcg <- "yes"
r1 <- predictRisk(e.cox, newdata = bissau1, time = 150)

## Compare the average risk across treatment groups:
c(mean(r0), mean(r1), mean(r1) - mean(r0))
```

```
[1]  0.04369355  0.02279141 -0.02090215
```

Standardization via riskRegression

```
e.ate <- ate(e.cox, treatment = "bcg", time = 150,
             data = bissau[bissau$agem %in% 0:2,])
summary(e.ate)
```

```
[...]
 - Difference in standardized risk (B-A) between time zero and '
risk(bcg=A) risk(bcg=B) difference           ci p.value
     0.0437      0.0228    -0.0209 [-0.03;-0.01] 0.00192
[...]
```

⚠ The uncertainty about the prediction should be accounted for.

Do not use:

```
unlist(t.test(r1, r0)[c("estimate","p.value")])
```

```
estimate.mean of x estimate.mean of y                p.value
        0.02279141         0.04369355         0.00000000
```

Take home message

The ATE enables to summarize complex treatment effects into a single number that is still interpretable

- machine learning technics can be used!
- more sophisticated estimators exist (double robust, TMLE)

The summarized effect is now population dependent:

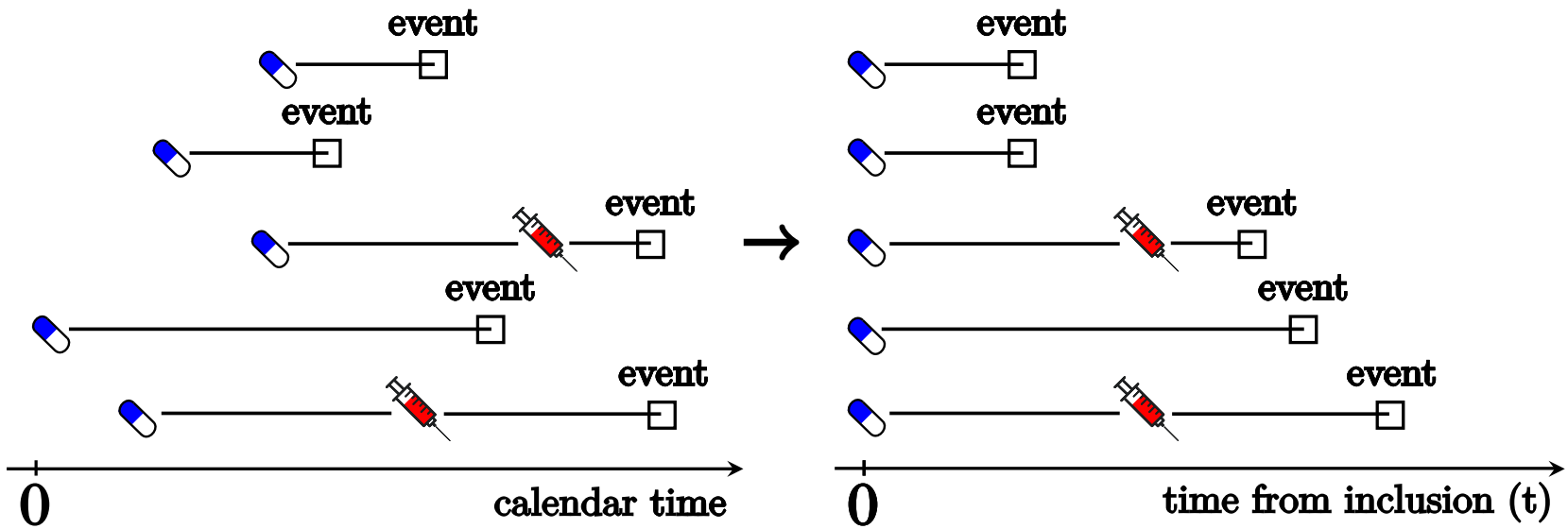
 should be performed over a representative population

$\rightarrow$ well suited for studies on national registries

Time varying exposures

Time varying exposures

Can you assess whether switching is beneficial? How?



Parameter of interest

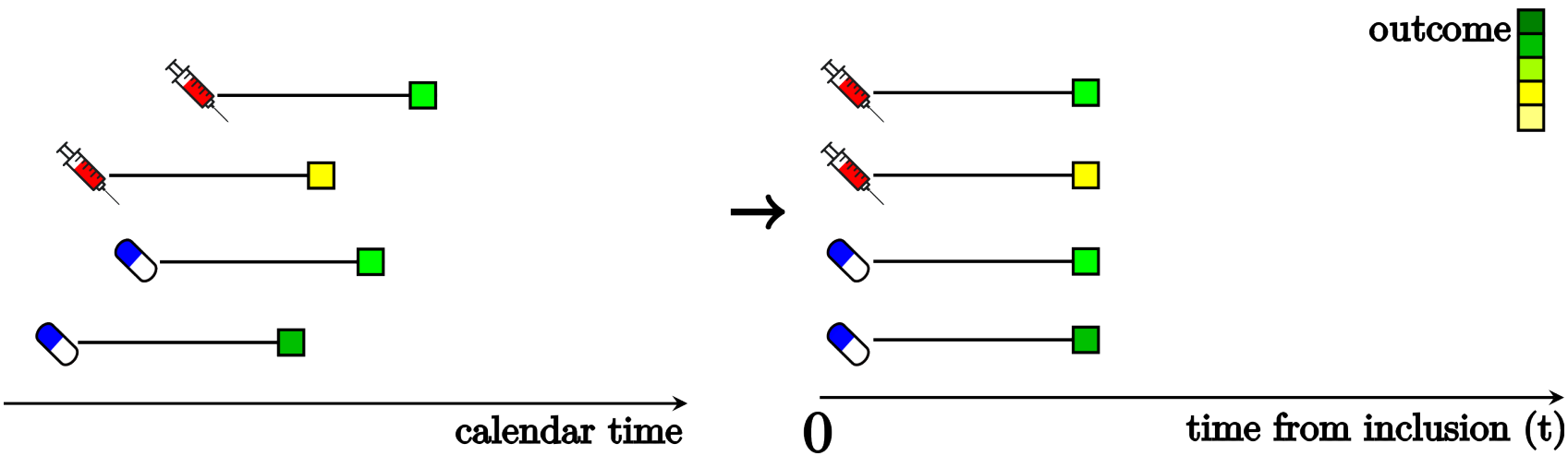
What do we mean by beneficial?

- **hazard**: the instantaneous risk of death is lower after switching compare to staying

- **risk**: unclear! (say at 1 year)
  - staying vs. switching after 1 month
  - staying vs. switching after 3 months
  - staying vs. switching if initial drug seems ineffective
  - staying vs. switching if initial drug seems harful

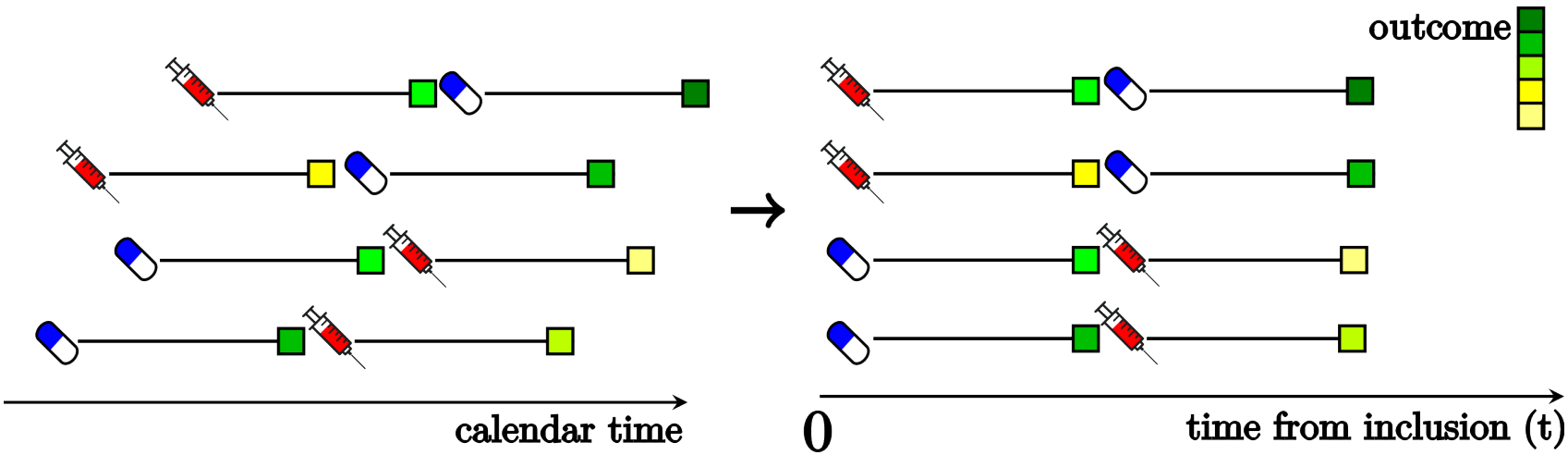
'Traditional' experimental studies

- single treatment received just after baseline



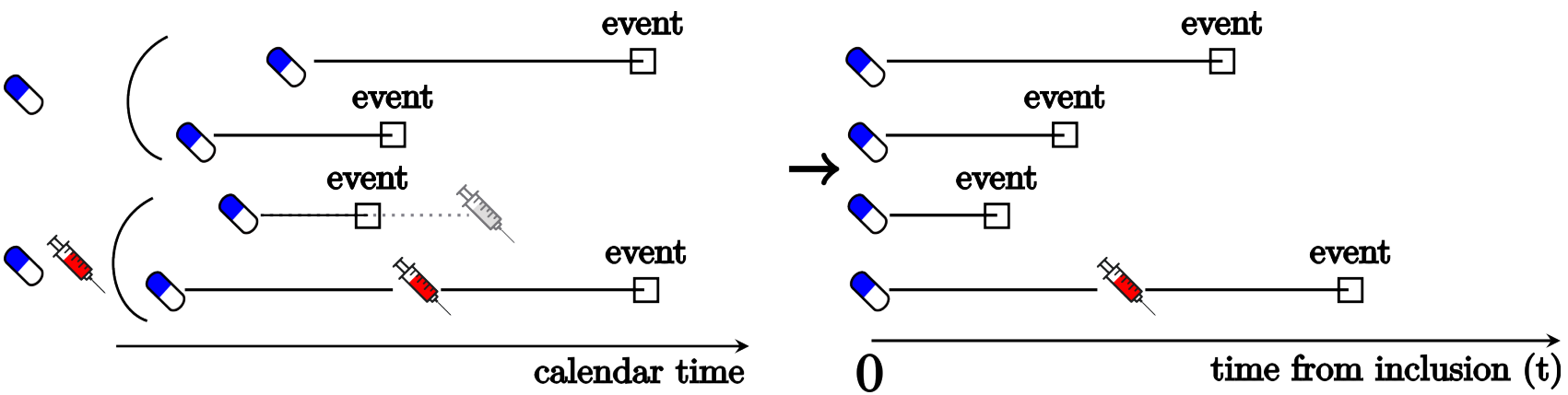
'Traditional' experimental studies

- single treatment received just after baseline
- cross over



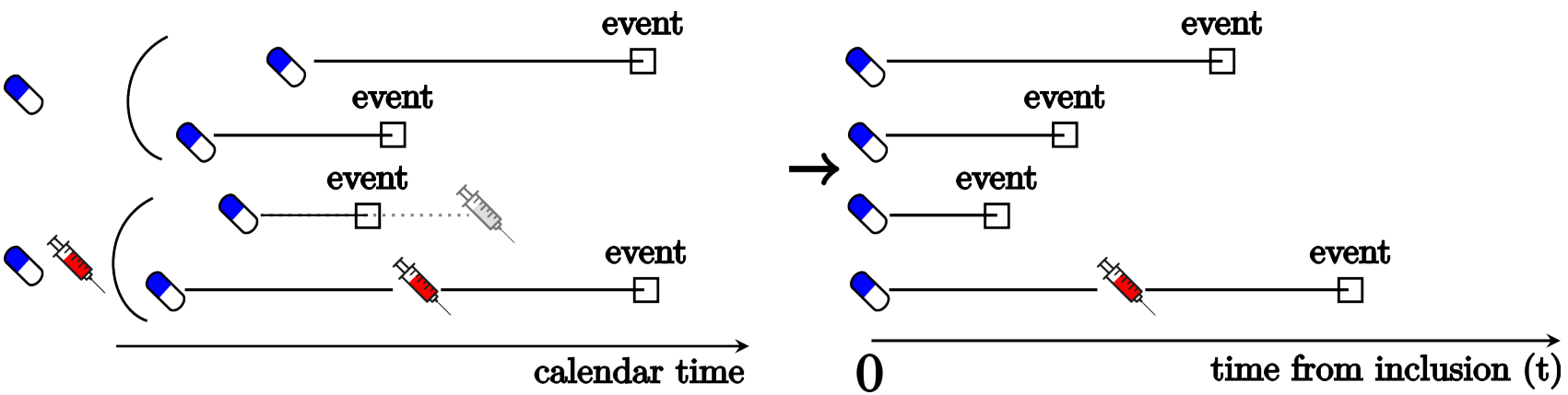
'Traditional' experimental studies

- single treatment received just after baseline
- cross over
- switch vs no switch between treatments



'Traditional' experimental studies

- single treatment received just after baseline
- cross over
- switch vs no switch between treatments



Immortal time bias: comparing patients who did not switch to those who did gives a survival advantage to those who switched. They 'cannot' die between inclusion and switch of treatment

Example 1 - (Lange and Keiding, 2014)

Letters to the Editor

Skin cancer as a marker of sun exposure

Brøndum-Jacobsen *et al.* recently published in this journal[1] analyses of Danish register data concerning myocardial infarction, hip fracture and death from any cause, using incidence of skin cancer as indicator of high exposure to sunlight. The basic idea in the paper is that those who get a skin cancer diagnosis at any age are supposed to have been more exposed to the sun during their life than those who do not, and apparently the authors find it relevant to use ordinary prospective survival analysis to compare incidence of myocardial infarction, hip fracture and death from any cause between the two groups: those who (at some point) get a skin cancer diagnosis and those who do not.

Example 1 - (Lange and Keiding, 2014)

Letters to the Editor

Skin cancer as a marker of sun exposure: a case of serious immortality bias

From Theis Lange* and Niels Keiding

Department of Biostatistics, Institute of Public Health, University of Copenhagen, Denmark

Brøndum-Jacobsen *et al.* recently published in this journal[1] analyses of Danish register data concerning myocardial infarction, hip fracture and death from any cause, using incidence of skin cancer as indicator of high exposure to sunlight. The basic idea in the paper is that those who get a skin cancer diagnosis at any age are supposed to have been more exposed to the sun during their life than those who do not, and apparently the authors find it relevant to use ordinary prospective survival analysis to compare incidence of myocardial infarction, hip fracture and death from any cause between the two groups: those who (at some point) get a skin cancer diagnosis and those who do not.

Unfortunately, such an analysis is seriously flawed, because the definition of one of the two groups to be compared conditions on the future: in order to get a skin cancer diagnosis, and thus become a member of the skin cancer group, it is at least necessary to survive until age of diagnosis, but the authors' analysis does not take this conditioning into account. Put another way: for those in the skin cancer group it is impossible to die until the age of diagnosis of the cancer, the so-called immortal person-time.[2]

It is seen in the lower left panel of Figure 2[1] that those who get non-melanoma skin cancer at some age have a hazard ratio of dying from any cause in the age interval 40–49 years of about 0.2 vs those who never get a non-melanoma skin cancer diagnosis. A main reason for this is probably that very few of those with non-melanoma skin cancer are at all at risk for dying—most of the members of this group get their skin cancer diagnosis at ages >50 years and are therefore by design immortal in the age interval 40–49.

Example 2 - (Shariff et al., 2008)

In the March 2007 issue of *JASN*, Hemmelgarn *et al.*[1] reported a 50% reduction in the risk for all-cause mortality for patients who had chronic kidney disease (CKD) and attended multidisciplinary care (MDC) clinics compared with those who received usual care. Their survival curves showed a clear divergence in rates of death between the two groups in the first 6 months of follow-up. We suggest that it is less plausible from a biologic perspective that use of MDC clinics immediately reduces the short-term risk for death. Rather, much of the early observed effect may be due to survivor treatment selection bias, also known as immortal time bias.

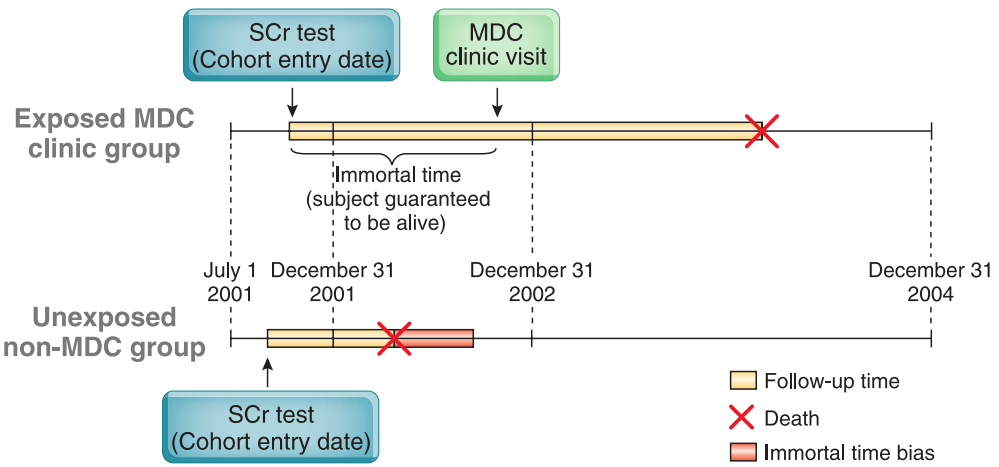


Figure 3. Immortal time bias. Situation in which MDC clinic visit occurred after serum creatinine test. Exposed patient was guaranteed to be alive between the test date and the clinic visit, resulting in a period of "immortal time."

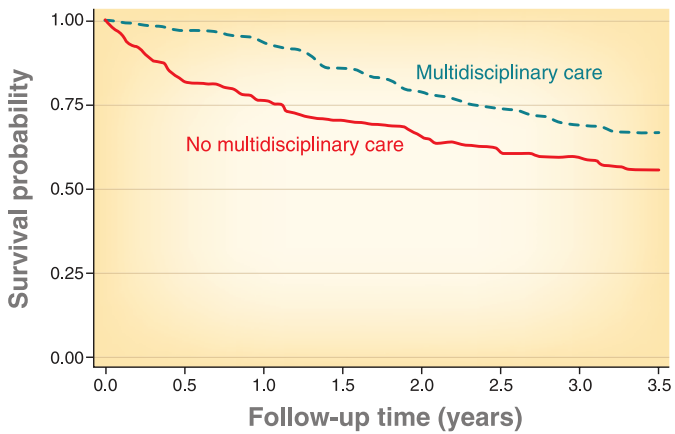


Figure 2. Kaplan-Meier survival curve

From person to person-time

We cannot distinguish switchers from non-switchers

id	event	start	stop	switch
1	TRUE	0	3.0	NA
2	TRUE	0	3.0	NA
3	TRUE	0	5.0	4.0
4	TRUE	0	6.0	4.5
5	TRUE	0	5.5	NA

Instead, we have at risk time before switch and after switch

id	event	start	stop	switch
1	TRUE	0.0	3.0	FALSE
2	TRUE	0.0	3.0	FALSE
3	FALSE	0.0	4.0	FALSE
3	TRUE	4.0	5.0	TRUE
4	FALSE	0.0	4.5	FALSE
4	TRUE	4.5	6.0	TRUE
5	TRUE	0.0	5.5	FALSE

Statistical analysis with time varying exposures

⚠ This is a difficult topic!

Cox model but requires strong assumptions:

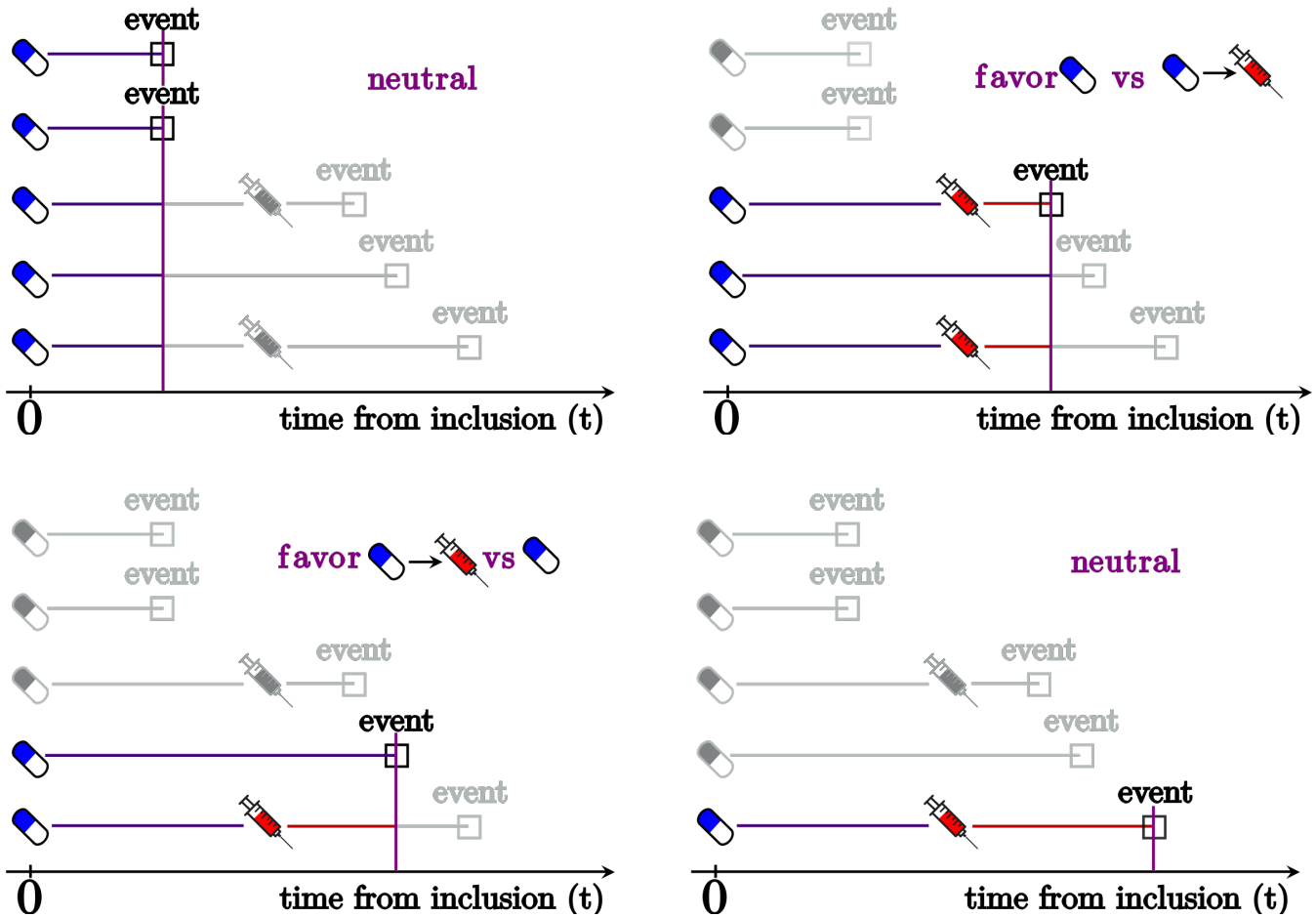
- reason for switching are not related to the outcome
- switching effect constant over time

```
e.cox <- coxph(Surv(start, stop, event) ~ switch,
               data = df.switch)
```

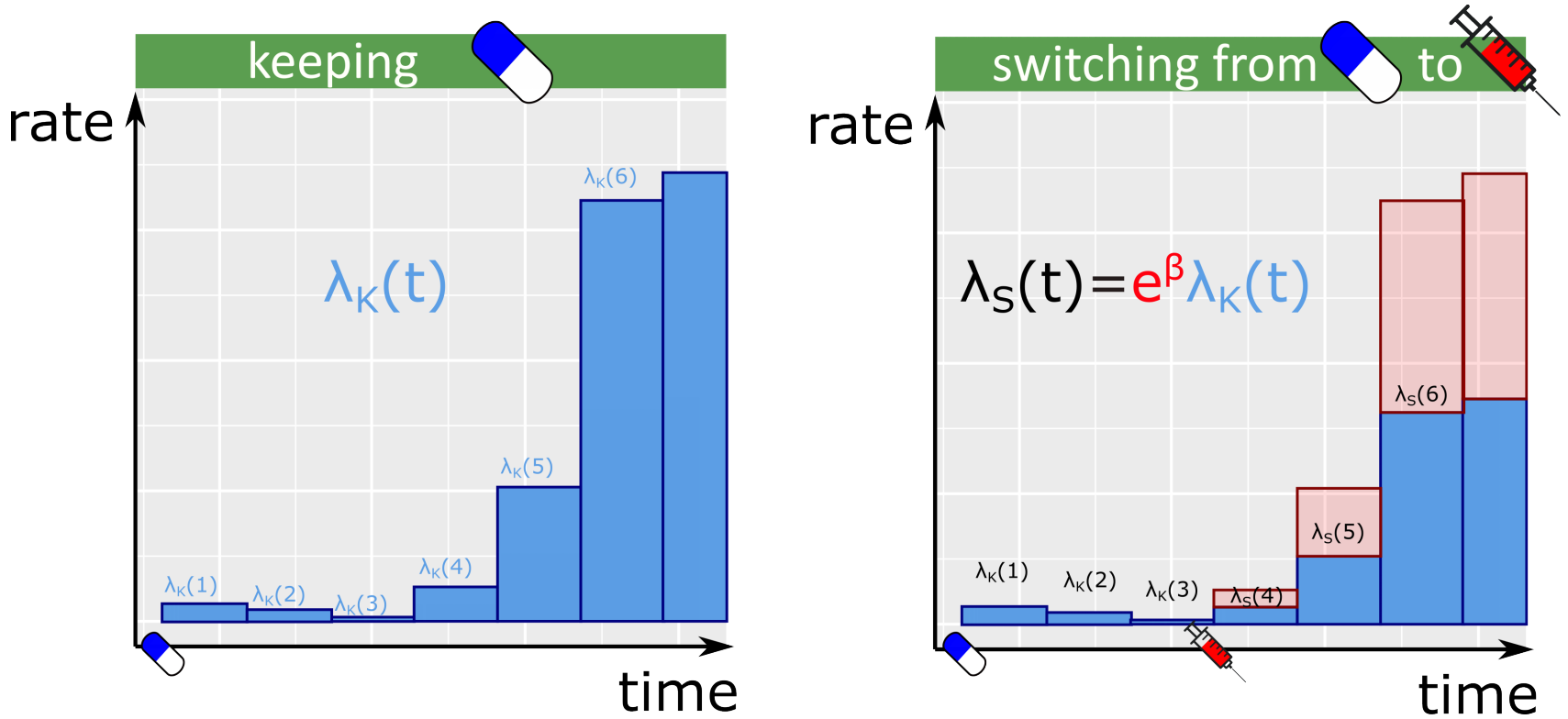
Otherwise more complex methods are needed (Hernán and Robins (2010), chapter 19-22), which involve modeling the probability of switching and using them to 're-weight' the data, hoping to rebalance confounders.

Intuition behind the Cox model

Matching: compare individuals at risk at the same time

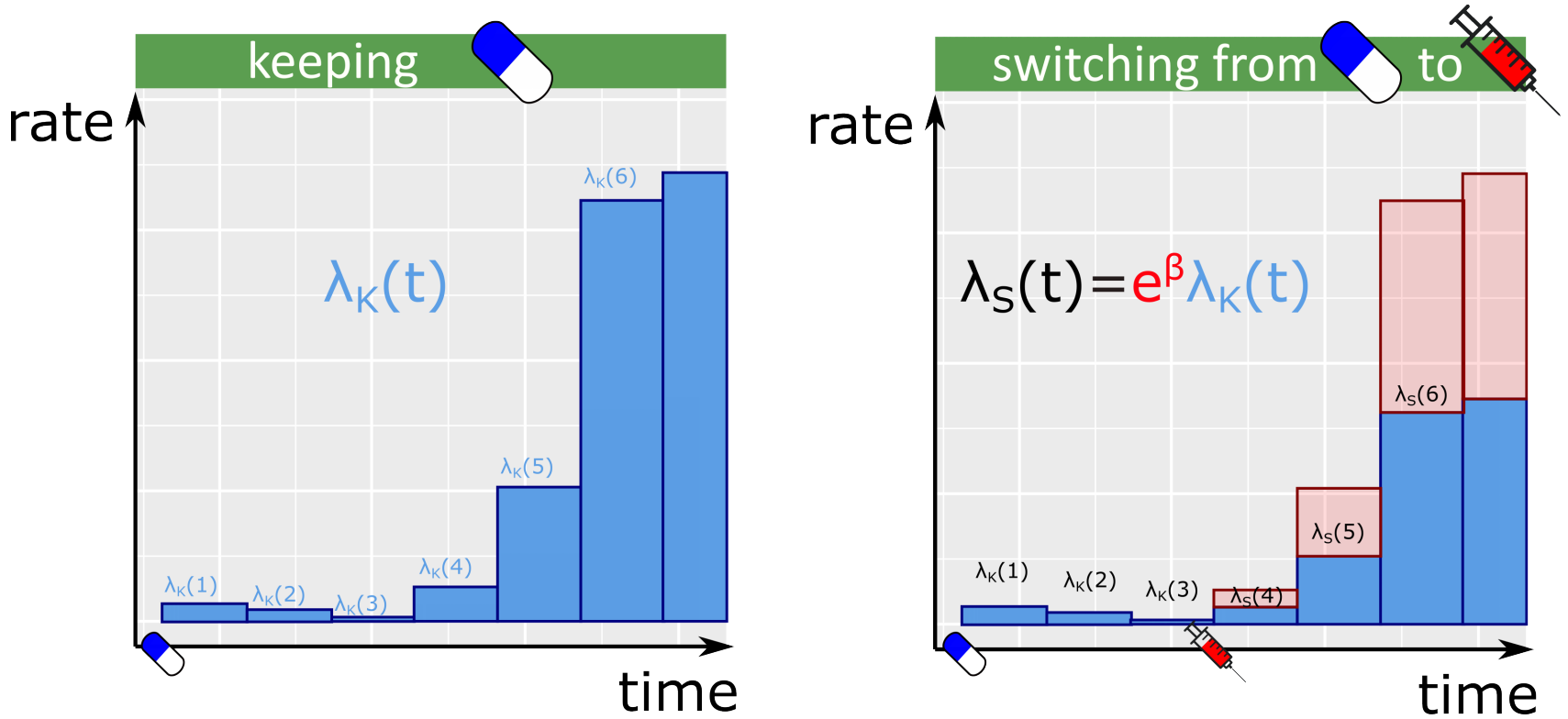


Representation of the Cox model



Multiplicative effect of the treatment (e^{β}) on the rates ($\lambda(t)$):

Representation of the Cox model

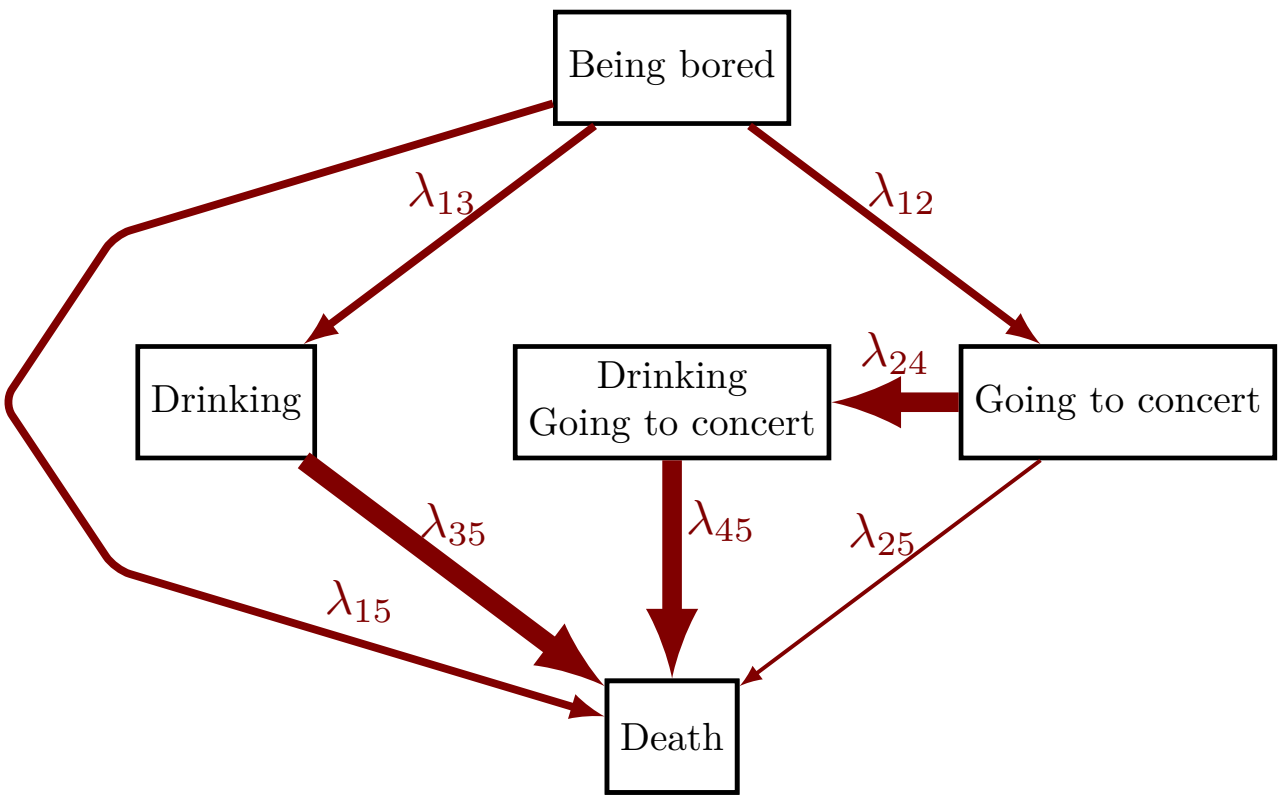


Multiplicative effect of the treatment (e^{β}) on the rates ($\lambda(t)$):

- same at all follow-up times
- same regardless to when the new treatment was initiated

Interpret carefully

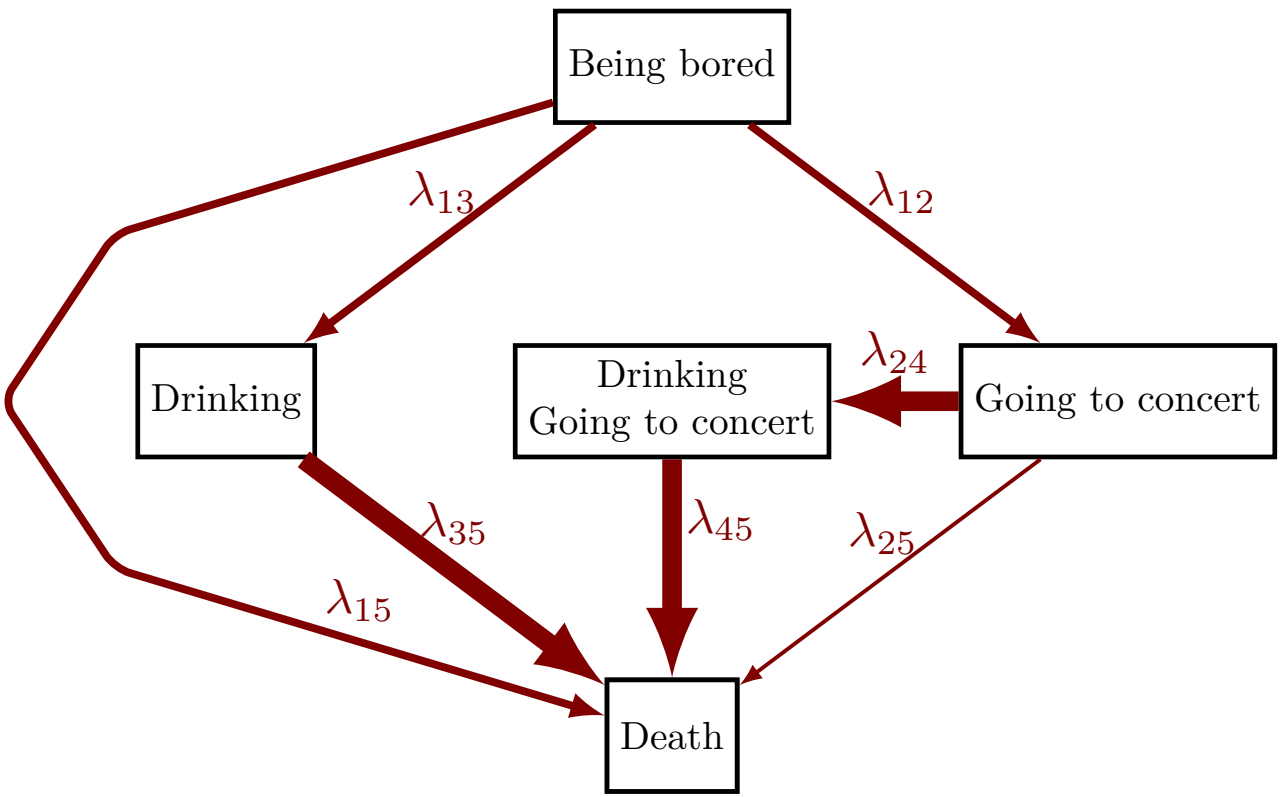
Going to concert vs. staying bored:



Interpret carefully

Going to concert vs. staying bored:

- lower *instantaneous* risk ($\frac{\lambda_{25}}{\lambda_{15}} < 1$)
- higher *long-term* risk (as one is likely to start drinking)



What we have seen today

- ✔ Illustration of the independent censoring assumption
  - Kaplan Meier as a re-weighting approach
  - Treating death as censoring is a bad idea
- ✔ Introduction to registry data
  - choice of the time scale
  - recognizing time varying exposure
  - dealing with individual specific follow-up times
- ✔ Standardization/ATE
  - summarize into single number the treatment effect (compatible with very flexible models)
  - positivity assumption
  - require a meaningful population
- ✔ Handling time varying exposures
  - what not to do: 'same as usual' $\rightarrow$ immortal time bias
  - what to do: split follow-up time
  - (greatly) complexify data analysis: reach for help

Reference I

Hernán, M. A., Alonso, A., and Logroscino, G. (2008). Commentary: Cigarette smoking and dementia: Potential selection bias in the elderly. *Epidemiology*, pages 448–450.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

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.

Lange, T. and Keiding, N. (2014). Skin cancer as a marker of sun exposure: a case of serious immortality bias. *International journal of epidemiology*, 43(3):971–971.

Shariff, S. Z., Cuerden, M. S., Jain, A. K., and Garg, A. X. (2008). The secret of immortal time bias in epidemiologic studies. *Journal of the American Society of Nephrology*, 19(5):841–843.

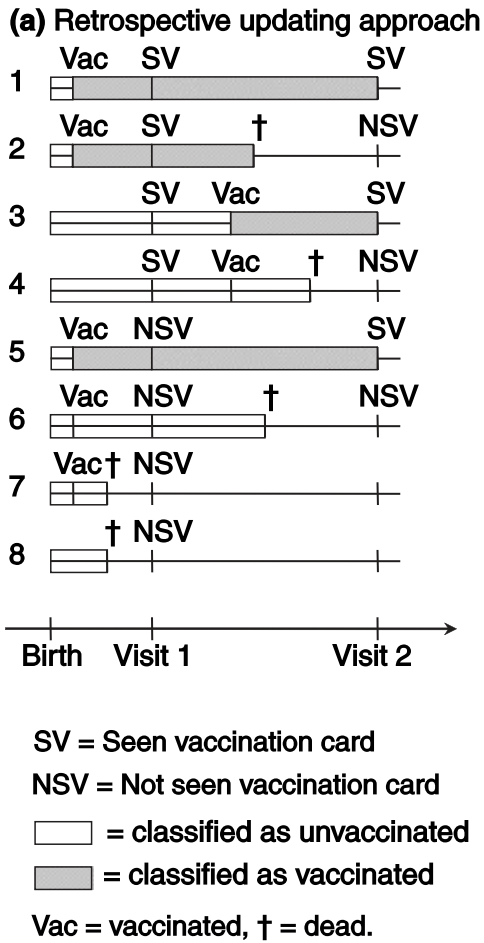
Survivorship bias

From Hernán et al. (2008):

> The article by Euser et al[1] in this issue of EPIDEMIOLOGY shows that study participants with complete follow-up are healthier and have better age-specific cognitive scores than those with incomplete follow-up. A well-known potential consequence of these differences is selection bias: when the analysis is restricted to individuals with complete follow-up (eg, those not too ill to participate), it is possible to find an exposure-outcome association that is not due to the causal effect of the exposure on the outcome.[2] An extreme case of "incomplete follow-up" for nonfatal outcomes is death; hence censoring by death may introduce selection bias. In studies of old people, this selection bias may be large because the death rate is high and death is often affected by the exposure.[3] Here we provide some empirical support for selection bias due to censoring by death in epidemiologic studies of the effect of cigarette smoking on risk of dementia.

Immortal time bias

From Jensen et al. (2007):

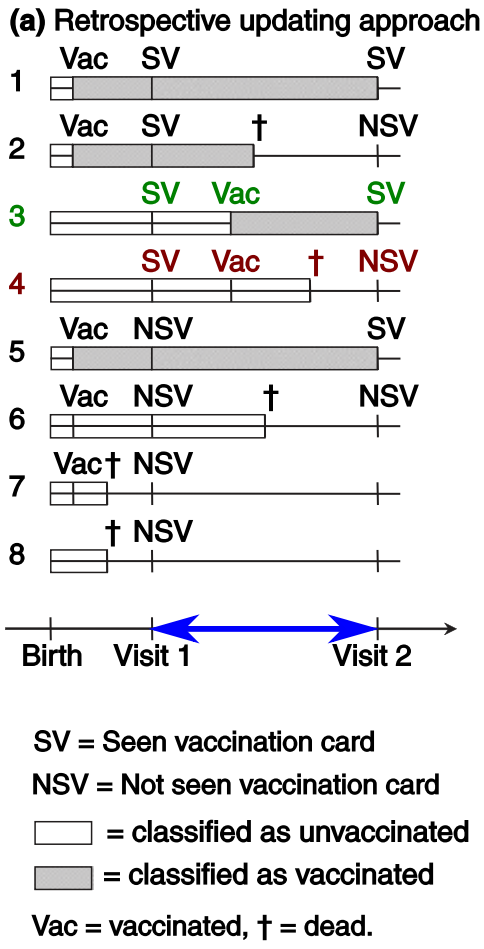


Retrospective updating approach

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

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.