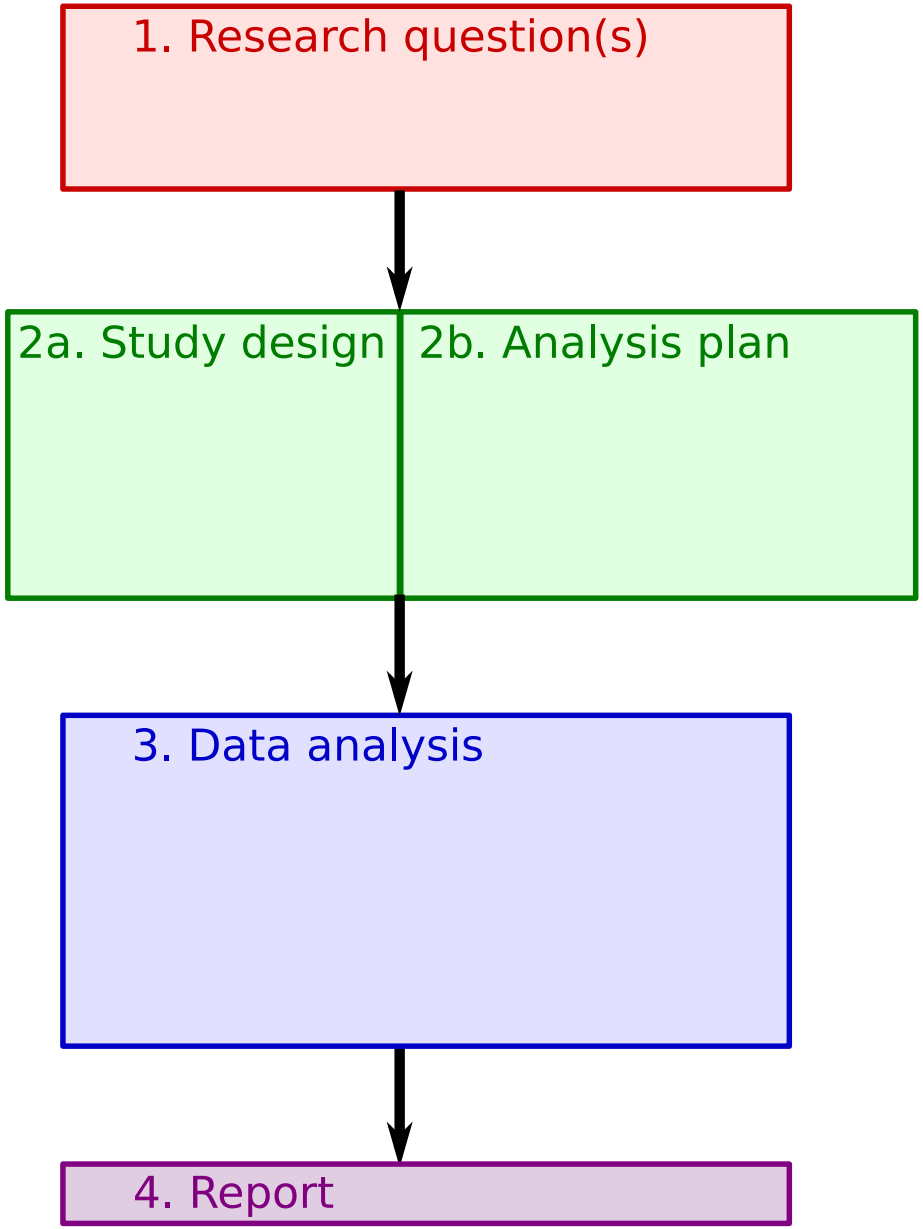
Reproducible science + 10 recommendations

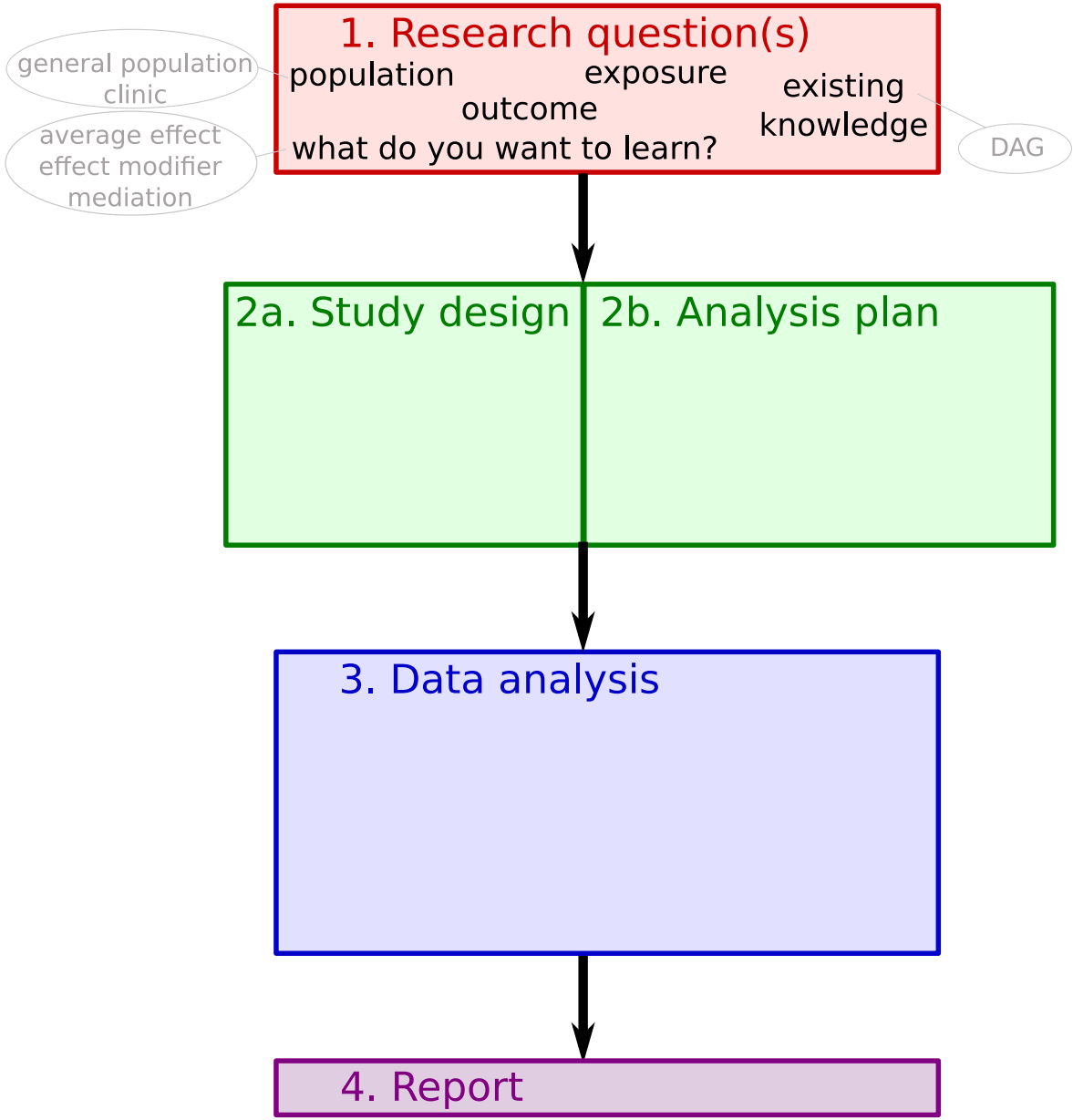
Brice Ozenne[1,2] - brice.mh.ozenne@gmail.com

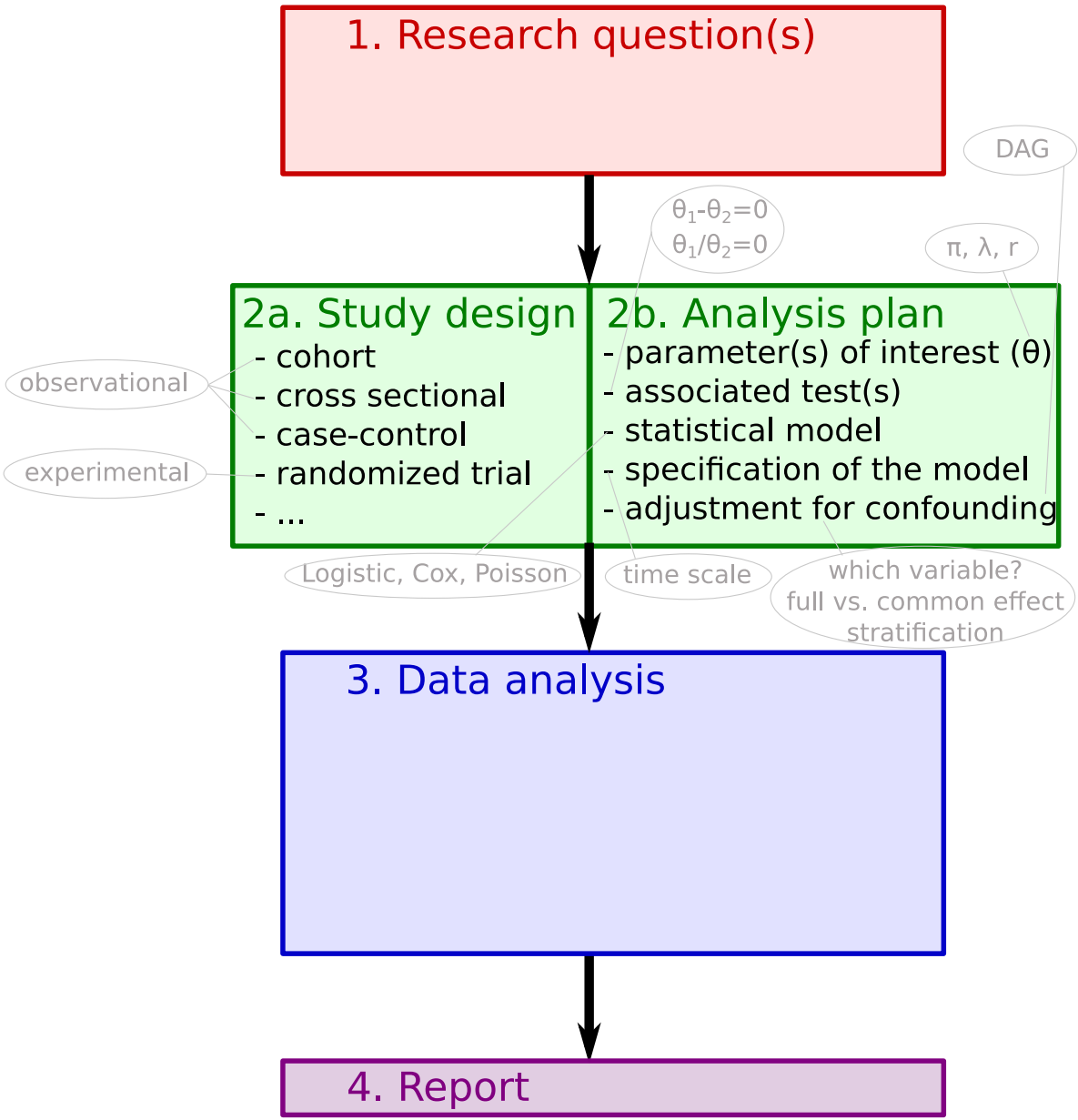
[1] Section of Biostatistics, Department of Public Health, University of Copenhagen

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

23 March 2023


1. Research question(s)
2a. Study design
2b. Analysis plan
3. Data analysis
4. Report


1. Research question(s)
population
exposure
outcome
existing knowledge
what do you want to learn?
general population
clinic
average effect
effect modifier
mediation
DAG
2a. Study design
2b. Analysis plan
3. Data analysis
4. Report


1. Research question(s)
DAG
θ_1-θ_2=0
θ_1/θ_2=0
π, λ, r
2a. Study design
- cohort
- cross sectional
- case-control
- randomized trial
- ...
2b. Analysis plan
- parameter(s) of interest (θ)
- associated test(s)
- statistical model
- specification of the model
- adjustment for confounding
observational
experimental
Logistic, Cox, Poisson
time scale
which variable?
full vs. common effect
stratification
3. Data analysis
4. Report

1. Research question(s)

population exposure existing knowledge

outcome

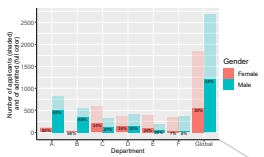
what do you want to learn?

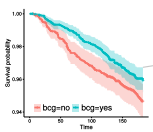
2a. Study design

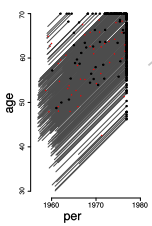
- cohort
- cross sectional
- case-control
- randomized trial
- ...

2b. Analysis plan

- parameter(s) of interest (θ)
- associated test(s)
- statistical model
- specification of the model
- adjustment for confounding

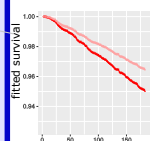






3. Data analysis

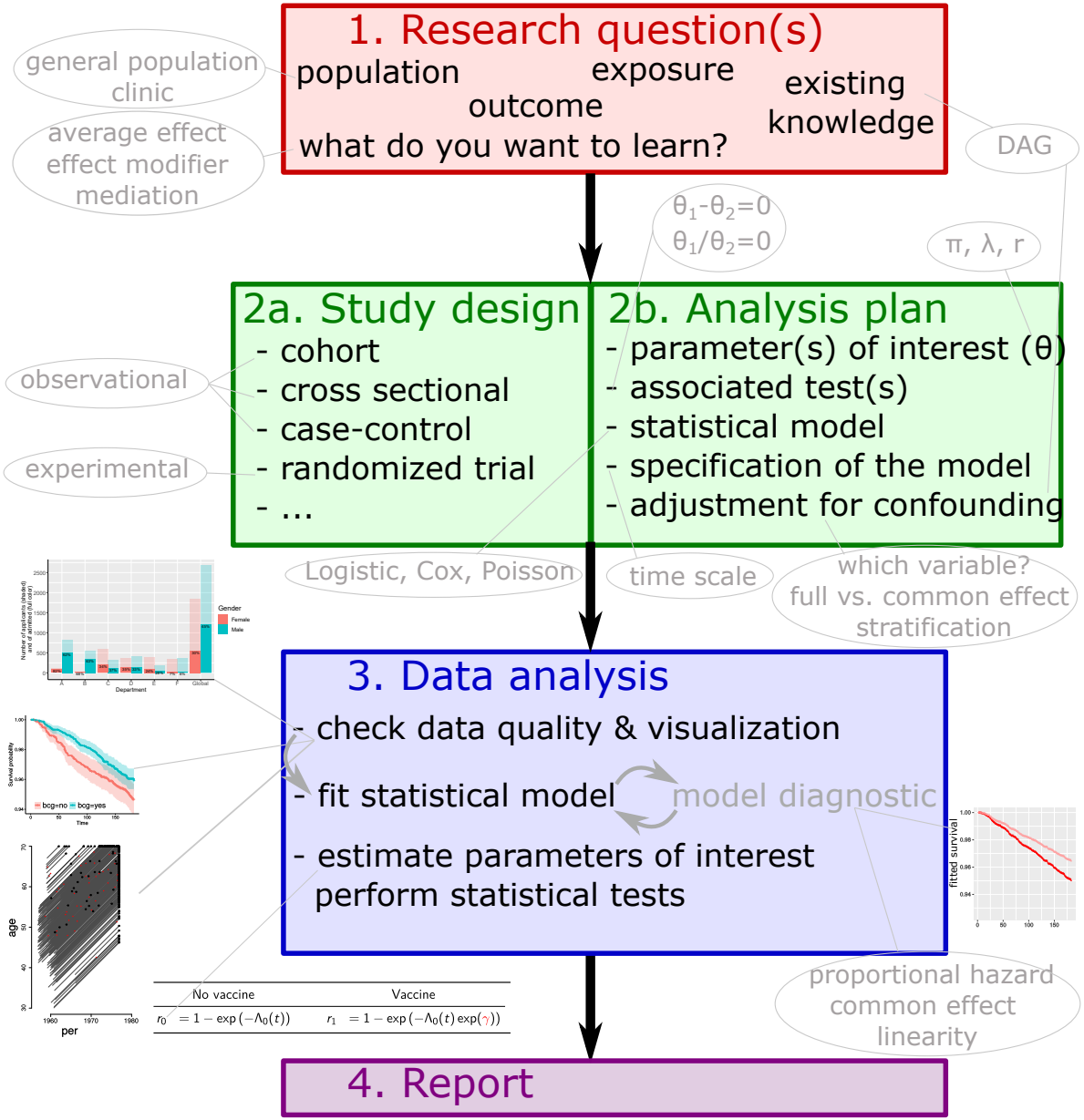
- check data quality & visualization
- fit statistical model ⇄ model diagnostic
- estimate parameters of interest
  perform statistical tests



No vaccine	Vaccine
$r_0 = 1 - \exp(-\Lambda_0(t))$	$r_1 = 1 - \exp(-\Lambda_0(t)\exp(\gamma))$

proportional hazard
common effect
linearity

4. Report


1. Research question(s)
population
exposure
outcome
existing knowledge
what do you want to learn?
general population
clinic
average effect
effect modifier
mediation
DAG
θ1-θ2=0
θ1/θ2=0
π, λ, r
2a. Study design
- cohort
- cross sectional
- case-control
- randomized trial
- ...
2b. Analysis plan
- parameter(s) of interest (θ)
- associated test(s)
- statistical model
- specification of the model
- adjustment for confounding
observational
experimental
Logistic, Cox, Poisson
time scale
which variable?
full vs. common effect
stratification
3. Data analysis
- check data quality & visualization
- fit statistical model
model diagnostic
- estimate parameters of interest
perform statistical tests
No vaccine
Vaccine
$r_0 = 1 - \exp(-\Lambda_0(t))$
$r_1 = 1 - \exp(-\Lambda_0(t)\exp(\gamma))$
proportional hazard
common effect
linearity
4. Report

Why making my analysis reproducible?

Why making my analysis reproducible?

For your future you:

- re-generate results/tables/graphs for review (months later)
- apply the same/similar approach on new projects (years later)

Why making my analysis reproducible?

For your future you:

- re-generate results/tables/graphs for review (months later)
- apply the same/similar approach on new projects (years later)

For collaborators:

- facilitate review by statistician/programmer
- explaining and sharing is facilitated

Why making my analysis reproducible?

For your future you:

- re-generate results/tables/graphs for review (months later)
- apply the same/similar approach on new projects (years later)

For collaborators:

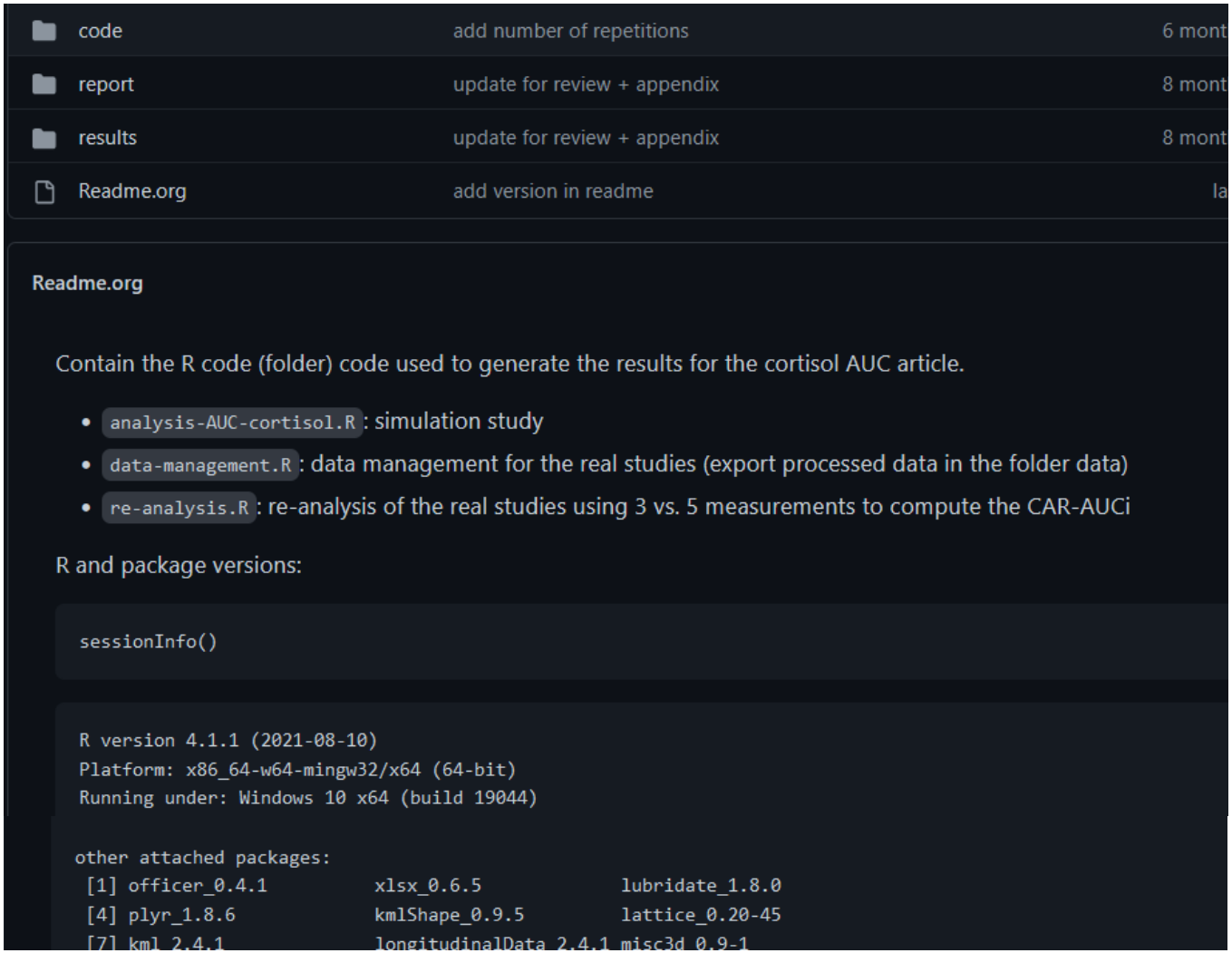
- facilitate review by statistician/programmer
- explaining and sharing is facilitated

For science:

- by 'cleaning' your code you may spot mistakes
- you provide easier access to your methodology

How to make my analysis reproducible

The are many ways . . . I'll illustrate mine on an example.



code	add number of repetitions	6 mont
report	update for review + appendix	8 mont
results	update for review + appendix	8 mont
Readme.org	add version in readme	la

Readme.org

Contain the R code (folder) code used to generate the results for the cortisol AUC article.

- `analysis-AUC-cortisol.R`: simulation study
- `data-management.R`: data management for the real studies (export processed data in the folder data)
- `re-analysis.R`: re-analysis of the real studies using 3 vs. 5 measurements to compute the CAR-AUCi

R and package versions:

```
sessionInfo()
```

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19044)

other attached packages:
 [1] officer_0.4.1          xlsx_0.6.5             lubridate_1.8.0
 [4] plyr_1.8.6             kmlShape_0.9.5         lattice_0.20-45
 [7] kml_2.4.1              longitudinalData_2.4.1 misc3d_0.9-1
```

Objective

Create a new folder containing:

- the data
- the programming instructions (R script, SAS script, . . .)
- specifics about the program used (version)

to reproduce results, tables, and figures of an article

- **only those but all of those**
- I usually do that at the end of a project when the structure of the article is 'stable'

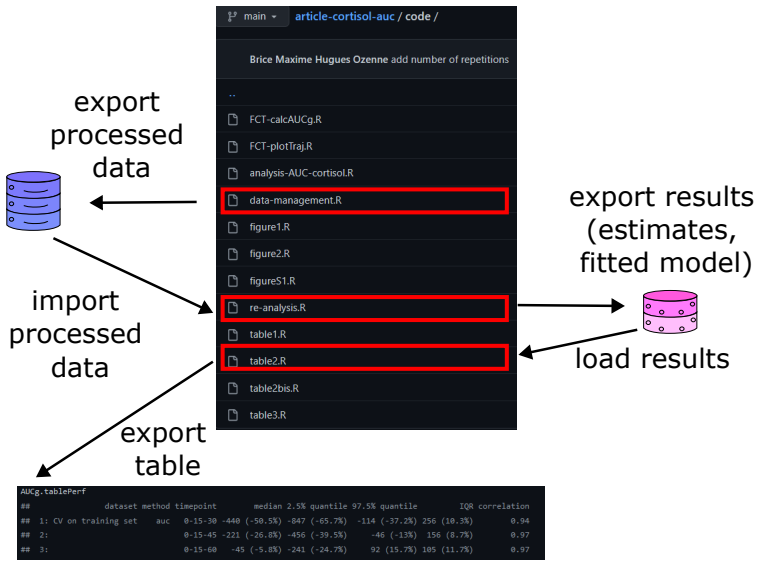
⚠ avoid unnecessary programming instruction (old analyses)!

⚠ if you share the folder consider carefully if you share the data or not

Programming instructions

I like to have:

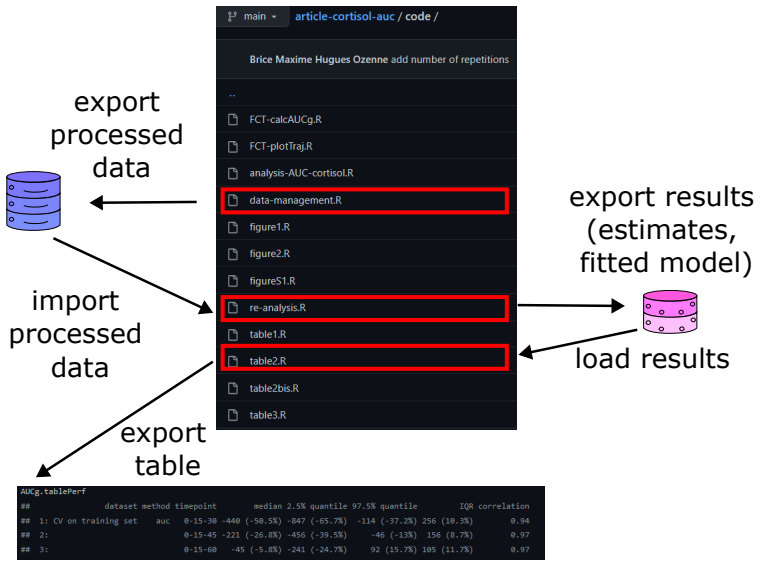
- a data management file
- data analysis files:
  one per research question
  or simulation/real data
- one file per table
- one file per figure



Programming instructions

I like to have:

- a data management file
- data analysis files:
  one per research question
  or simulation/real data
- one file per table
- one file per figure



Saving intermediate results:

- save time when long analyses (avoid recomputation)

Relation to the method section (1/2)

Reading code is often harder than reading english

- people use different programming languages
- require familiarity with the tools used

The method section should (still) fully describe:

- data manipulation during data management (exclusion, transformation, imputation, . . .)
- statistical models and tests

Check the correspondence between your (clean) code and what you describe!

Relation to the method section (2/2)

Saying that your analysis has been done in R 2.4.0 does not help much for reproducing the results.

Sharing your code (and if possible data), with a description of your software version (e.g. `sessionInfo`) is useful.

Relation to the method section (2/2)

Saying that your analysis has been done in R 2.4.0 does not help much for reproducing the results.

Sharing your code (and if possible data), with a description of your software version (e.g. `sessionInfo`) is useful.

Side note: "A Cox model was used to relate the survival time of patients to their vaccination and age group." miss one key information

- what is the parameter of interest: hazard ratio, difference in 1 year risk, ...

Challenges

Making the analysis reproducible is not really time consuming

- you spend some time now but save time later. It is mostly copy pasting.
- it is a good sanity check

Making the code readable can be time consuming

- commenting the code, having meaningful names, appropriate spacing, avoid long lines [1]

⚠ bugs can be created by renaming, or modifying lines of code. So better to start clean than to fix afterwards.

[1] example of coding style `https://style.tidyverse.org/`

Some recommandations

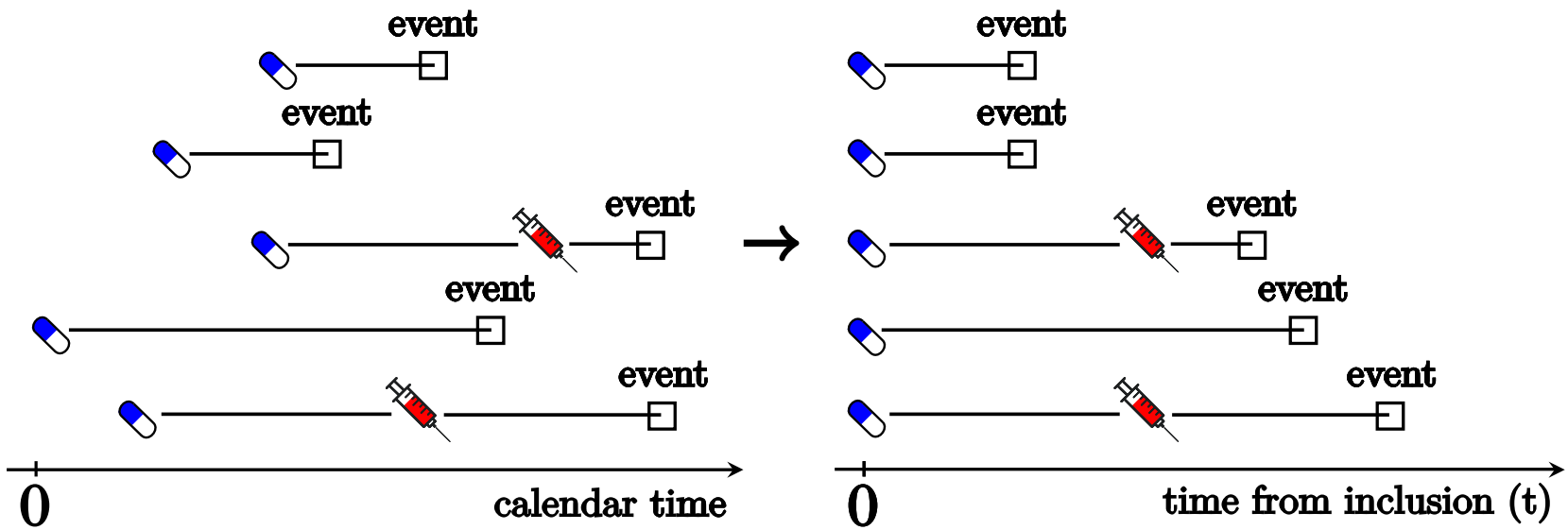
1. Start with a plan

- First define the research questions(s)
- Then look at the data

Why?

- save time: avoid to get lost in the possibilites
- validity: keep track of how many "tests" have been performed
- avoid temptation: hard to think "independently" after having looked at the data

2. Respect time-varying exposures



Immortal time bias when comparing never-switchers to switchers.

3. Use matching with care

Matching can be used to:

- (try to) adjust for unobserved confounders in an observational study.
- efficiently sample based on important factors

Data analysis must take account of the matching:

- more complex to understand and carry out
  (e.g. individual matching: conditional logistic regression)

⚠ some coefficients cannot be interpreted
(e.g. frequence matching on age: age coefficient)

- not always clear what one really adjust on

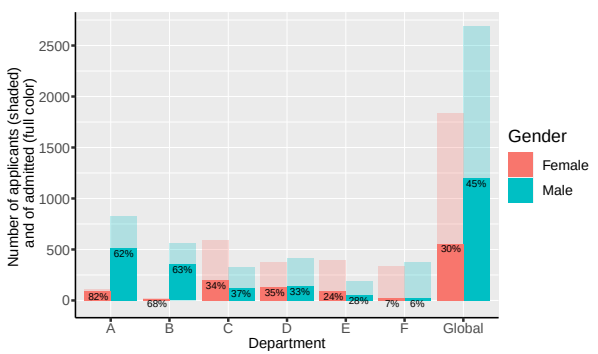
  ⚠ risk of overmatching
  e.g. twin study relating diet to obesity

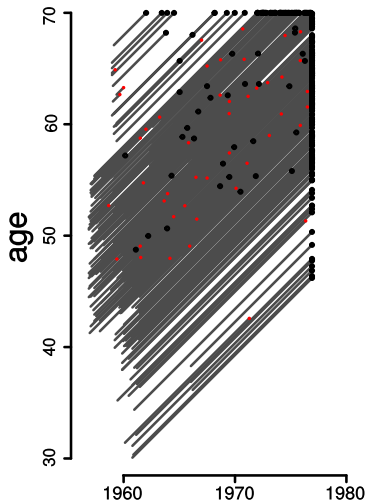
4. Inspect your data before fitting models

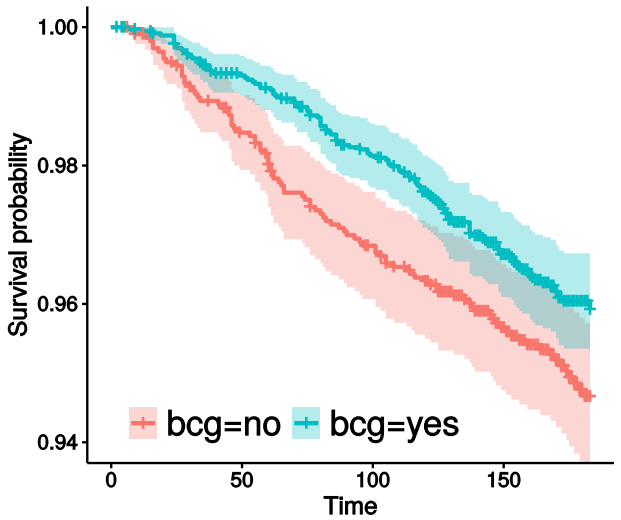
Table:

	set													exposure	
	1	3	4	5	7	8	9	10	11	12	14	16	17	0	1
control	2	2	1	3	3	2	3	2	3	3	3	3	3	46	0
case	1	1	1	1	1	1	1	0	1	1	1	1	1	15	0

Graph: Histogram, Lexis diagram, Kaplan Meier curves, ...







5. Be concerned of what you don't observe

Drop-out / censoring: (I could have observed it)

- completely at random?
- related to the outcome? to the exposure?

→ investigate possible cause of drop-out/missingness?

Competing risks: (it cannot happen anymore)

- Specialized tools (e.g. cause specific methods)

⚠ the treatment should not reduce the risk of a disease by killing people

6. Model specification matters (1/2)

The name of a model often matters less than how you use it:

```
## 1
glm(event ~ exposure + age,
    family = poisson, offset = log(time), data = dSplit)
## 2
glm(event ~ period + exposure + age,
    family = poisson, offset = log(time), data = dSplit)
## 2.bis
coxph(Surv(event, time) ~ exposure + age,
      data = d)
## 3
glm(event ~ period * exposure + age,
    family = poisson, offset = log(time), data = dSplit)
```

- "We used a Poisson model to . . . " is a rather vague statement

6. Model specification matters (2/2)

When fitting a Poisson regression . . .

. . . remember the offset

- and the log-transform

. . . remember to specify the family argument

- otherwise `glm` will perform a "standard linear regression"

. . . remember to properly model the time effect (Lexis macro)

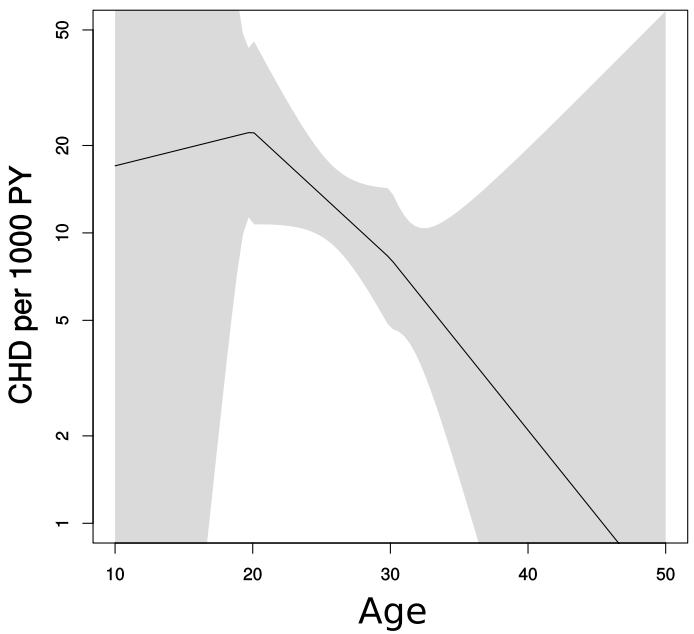
- otherwise too unflexible model

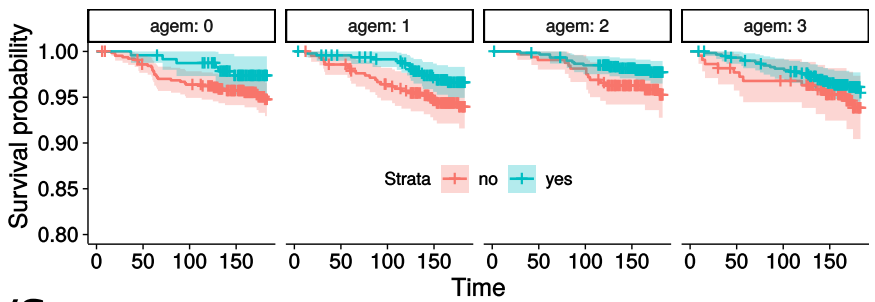
7. Vizualize the fitted model

In each strata or for specific individuals:

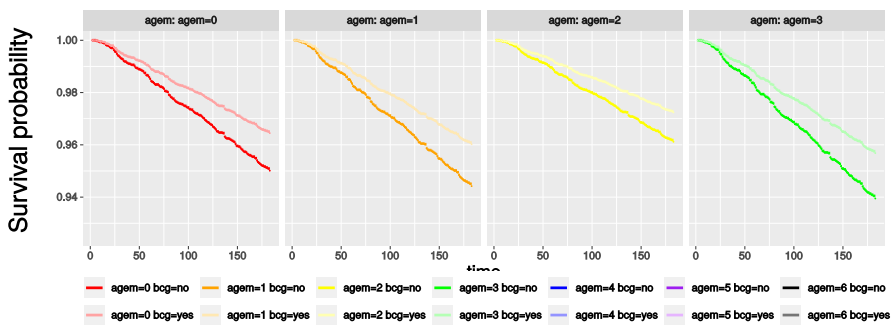
- Logisitic model: fitted prevalence/risk
- Poisson model: fitted rate/survival
- Cox model: fitted survival

Is that what you "wanted"? Does that match the observed data?





VS.



8. Don't get trapped by the software

When possible, do examples "by hand"

- softwares do calculation for you (+,-,*,/), not magic!

Labels of regression coefficients may be misleading

- in presence of interactions

```
             coef exp(coef) se(coef)      z Pr(>|z|)
bcg      -0.56991   0.56557  0.20445 -2.788  0.00531 **
dtp       0.17430   1.19041  0.72170  0.242  0.80916
bcg:dtp   0.21183   1.23594  0.74364  0.285  0.77576
```

Regression coefficients are tools:

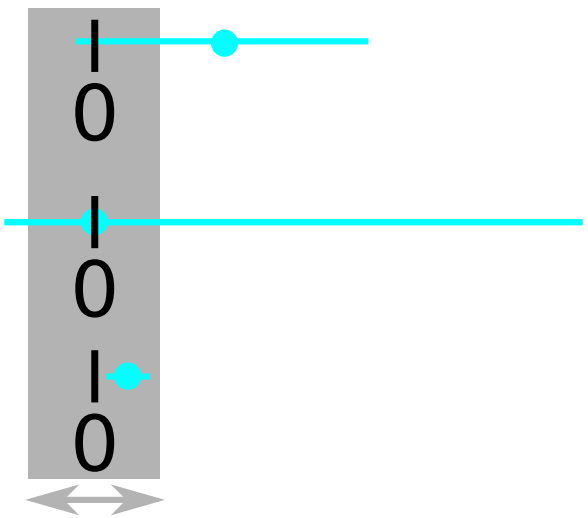
- not (necessarily) parameters of interest

```
                          Estimate Std. Error      z Pr(>|z|)
 exp(bcg)                 0.565582   0.115605 -2.788   0.0157 *
 exp(dtp)                 1.190413   0.859121  0.242   0.9926
 exp(bcg + dtp + bcg:dtp) 0.832102   0.145368 -1.052   0.6362
(Adjusted p values reported -- single-step method)
```

9. Absence of evidence is not evidence of absence

To "accept" the null hypothesis:

- look at the confidence intervals
- not at p-values!



cannot tell

cannot tell

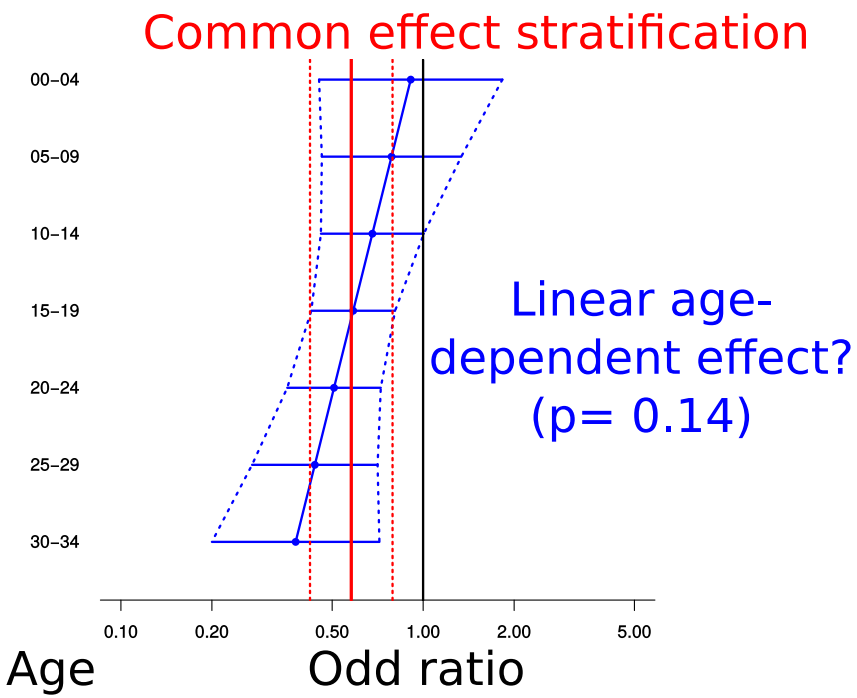
(statistically) significant
(clinically) irrelevant

τ Minimal clinically important difference

9. Absence of evidence is not evidence of absence

To "accept" the null hypothesis:

- look at the confidence intervals
- not at p-values!



10. Know when to ask for help

Not too early:

- you know what you want

Not too late:

- you are keen on implementing suggestions!

Try to write down as precisely as possible:

- your question
- what you have tried to address it

Statistical advisory service at KU:

- over the phone consultancy
- 20 minute meeting
- statistical café