Statistical View on Reproducible Science

Brice Ozenne[1,2]

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

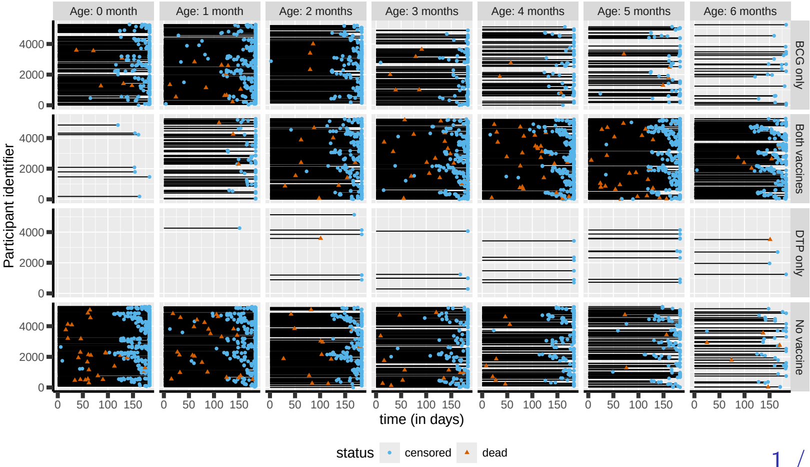
[2] Section of Biostatistics, Department of Public Health, University of Copenhagen.

02/27 - Dept. Of Oncology's Research Day 2026

Case study: Guinea-Bissau study (Kristensen et al., 2000)

Observational study following 5274 babies during 6 months

- two possible vaccines: BCG, DTP



Possible statistical analysis

Main research question

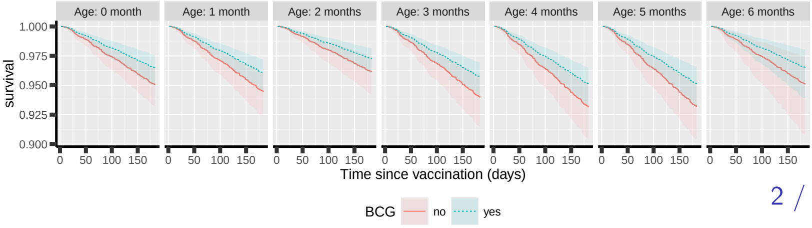
- assessing the effect of BCG on 6-month survival.

Statistical model:

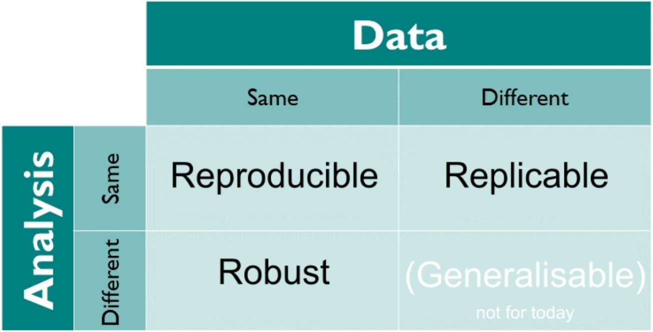
- Cox adjusted for age as categorical variable
  → proportional hazard assumption
  → independent censoring conditional on age and BCG
  → ignore other vaccines

Results: average difference in survival

- 0.0156 [0.003;0.029] (p=0.018)



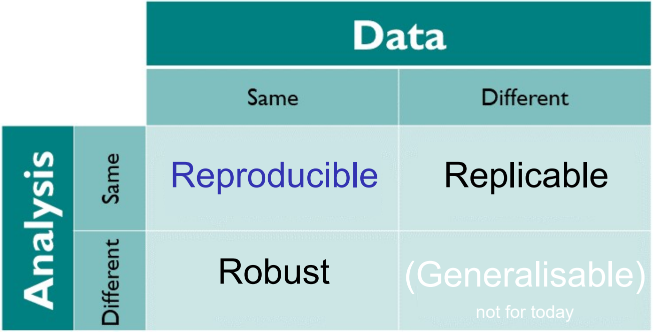
What are we aiming at?



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	(Generalisable) not for today

https://book.the-turing-way.org/reproducible-research/overview/overview-definitions/

What are we aiming at?



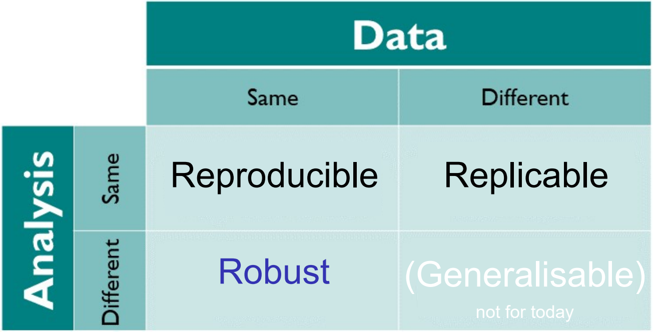
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	(Generalisable) not for today

https://book.the-turing-way.org/reproducible-research/overview/overview-definitions/

Reproducibility: precisely the same results when re-running the software

- I see that as a **sanity check**

What are we aiming at?



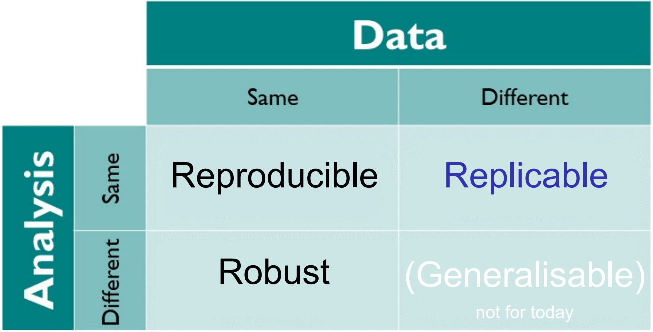
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	(Generalisable) not for today

https://book.the-turing-way.org/reproducible-research/overview/overview-definitions/

Robustness: comparable results when varying modeling assumptions.

- I see that as a **sensitivity analysis**

What are we aiming at?



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	(Generalisable) not for today

https://book.the-turing-way.org/reproducible-research/overview/overview-definitions/

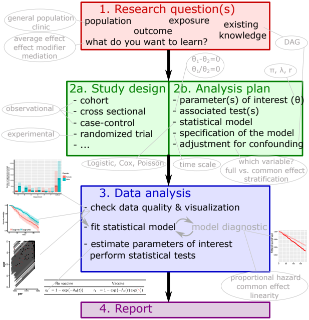
Replicability: comparable results with another dataset.

- I see that as a **sensitivity analysis**
- closely related to the **frequentist properties** of the analysis WHEN the new dataset is based on a similar population.

It can be surprisingly challenging! (1/3)

Reproducibility: possible obstacles

- **Time**: added work at the end of a (long) process
- **Complexity**:
  - many (sub)-cohorts
  - different software that may evolve
  - time-consuming analyses
- **Organisation**: project involving multiple researchers, some may have left the project with their skills.



1. Research question(s)

population exposure outcome existing knowledge

what do you want to learn?

general population
clinic

average effect
effect modifier
mediation

DAG

θ_1-θ_2=0
θ_1/θ_2=0

π, λ, r

2a. Study design
- cohort
- cross sectional
- case-control
- randomized trial
- ...

observational

experimental

2b. Analysis plan
- parameter(s) of interest (θ)
- associated test(s)
- statistical model
- specification of the model
- adjustment for confounding

Logistic, Cox, Poisson

time scale

which variable?
full vs. common effect
stratification

3. Data analysis
- check data quality & visualization
- fit statistical model
- estimate parameters of interest
perform statistical tests

model diagnostic

proportional hazard
common effect
linearity

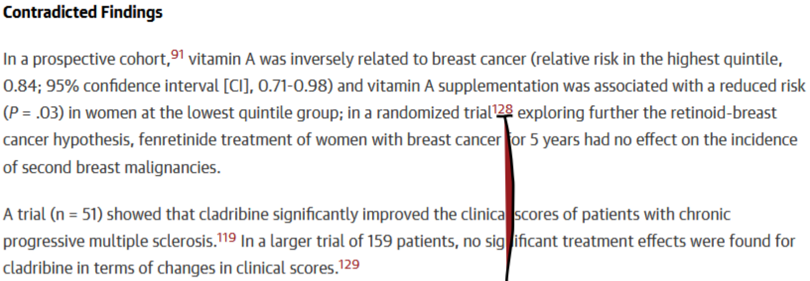
age
per
fitted survival
No vaccine
Vaccine
$r_0 = 1 - \exp(-\Lambda_0(t))$
$r_1 = 1 - \exp(-\Lambda_0(t)\exp(\gamma))$

4. Report

It can be surprisingly challenging! (2/3)

Replicability: how to define 'comparable results'?

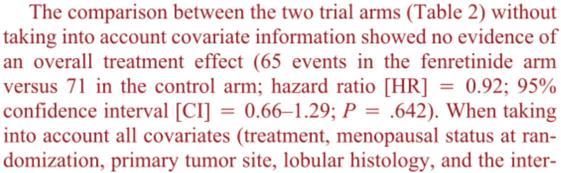
What do you think of this paragraph? (Ioannidis, 2005)



Contradicted Findings

In a prospective cohort,[91] vitamin A was inversely related to breast cancer (relative risk in the highest quintile, 0.84; 95% confidence interval [CI], 0.71-0.98) and vitamin A supplementation was associated with a reduced risk (*P* = .03) in women at the lowest quintile group; in a randomized trial[128] exploring further the retinoid-breast cancer hypothesis, fenretinide treatment of women with breast cancer for 5 years had no effect on the incidence of second breast malignancies.

A trial (n = 51) showed that cladribine significantly improved the clinical scores of patients with chronic progressive multiple sclerosis.[119] In a larger trial of 159 patients, no significant treatment effects were found for cladribine in terms of changes in clinical scores.[129]



The comparison between the two trial arms (Table 2) without taking into account covariate information showed no evidence of an overall treatment effect (65 events in the fenretinide arm versus 71 in the control arm; hazard ratio [HR] = 0.92; 95% confidence interval [CI] = 0.66–1.29; *P* = .642). When taking into account all covariates (treatment, menopausal status at randomization, primary tumor site, lobular histology, and the inter-

It can be surprisingly challenging! (2/3)

Replicability: how to define 'comparable results'?

- "Unfortunately, many popular and readily accessible methods for ascertaining replicability, such as comparing significance levels across studies or eyeballing confidence intervals, are generally ill suited to the task of comparing results across studies." (Spence and Stanley, 2024)
- Does it even make sense to compare estimates relative to different estimands?
  (here relative risk vs. hazard ratio)

It can be surprisingly challenging! (3/3)

Robustness: how to define 'comparable results'?

Sensitivity analyses are typically

- using less restrictive assumptions
- considering subset of the dataset

It can be surprisingly challenging! (3/3)

Robustness: how to define 'comparable results'?

Sensitivity analyses are typically not expected to lead to the same p-values when they consist in:

- using less restrictive assumptions
- considering subset of the dataset

(higher p-values are expected when the original model was correct)

It can be surprisingly challenging! (3/3)

Robustness: how to define 'comparable results'?

Sensitivity analyses are typically not expected to lead to the same p-values when they consist in:

- using less restrictive assumptions
- considering subset of the dataset

(higher p-values are expected when the original model was correct)

"the odds ratio or hazard ratio comparing treated and untreated individuals **will change** upon including a baseline covariate in the model, whenever that covariate is associated with the outcome" (Daniel et al., 2021)

- no matter how large the sample size
- even when there is no confounding

Reproducibility

(same results when re-running the software)

- principles

- illustration on a simplified project

Why 'bother' making my analysis reproducible?

Avoid embarrassing situations to your future self:

Facilitating collaborations:

Moral responsibility:

Why 'bother' making my analysis reproducible?

Avoid embarrassing situations to your future self:

- by cleaning-up your code, you may spot mistakes
- re-generate results/tables/graphs for review (months later)
- explain apparently conflicting results to other researchers working on the same data (months to years later)

Facilitating collaborations:

- implement last minute feedback about the analysis/plots
- recycling your code for other projects

Moral responsibility:

- makes it possible for co-authors/reviewers to spot mistakes
- transparency: provide all details about the analysis

Personal opinion

Ensuring reproducibility is like keep the kitchen clean:

- 1. you can make a mess and only clean at the end, hoping to save time and getting help from cleaning machines.
- 2. you can think about what to cook in which order and clean after you are done with each dish. The final cleaning should not be hard to do.

Personal opinion

Ensuring reproducibility is like keep the kitchen clean:

- 1. you can make a mess and only clean at the end, hoping to save time and getting help from cleaning machines.
- 2. you can think about what to cook in which order and clean after you are done with each dish. The final cleaning should not be hard to do.

I prefer option 2, trying to be organised **and** 'low tech'.

- others statisticians like to use the R package targets
- I have not tried using AI to ensure/check reproducibility.

Personal opinion

Ensuring reproducibility is like keep the kitchen clean:

- 1. you can make a mess and only clean at the end, hoping to save time and getting help from cleaning machines.
- 2. you can think about what to cook in which order and clean after you are done with each dish. The final cleaning should not be hard to do.

I prefer option 2, trying to be organised **and** 'low tech'.

- others statisticians like to use the R package targets
- I have not tried using AI to ensure/check reproducibility.

Option 2 is illustrated in a Github repository (link) with a 'template project' inspired from the Guinea-Bissau study

Organizing your folder

Documentation is time consuming - and not very stimulating.

- file organisation and clever naming can mitigate the need for documentation

Be principled and consistent, e.g.

- 1 file per 'task': data-processing, main analysis, secondary analysis, create table 1, create figure 1
- file `figure1.R` generates figure 1 in the article

⚠ easier at the article stage, when it has been decided what the figures and tables should be

⚠ files name should be understandable by humans and by machines (avoid special characters)

How to document the software?

People sometimes write in the method section of their paper:

- "All statistical analyses were performed using R (version 4.0.2, R Foundation for Statistical Computing, Vienna, Austria)."

What does this achieve?

How to document the software?

People sometimes write in the method section of their paper:

- "All statistical analyses were performed using R (version 4.0.2, R Foundation for Statistical Computing, Vienna, Austria)."

What does this achieve?

- give credit to the R software (no need to report the version! Use `citation()`)
- ~~help with reproducibility/replicability~~

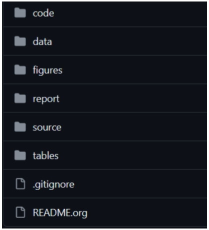
Details about software implementation should be reported separately as they are usually pretty lengthy:

"The R software (R Core Team, 2022) was use to implement the statistical analysis. The source code, the version of R and of related software package, can be found at `https://github.com/bozenne/article-template/tree/main`."

How I work - during the project

Create a new folder containing sub-folders:

- `code`: R code
- `data`: processed data
- `source`: original data
- `report`, `figure`, `tables`: output of the analysis

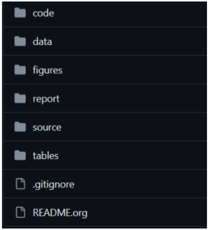


code
data
figures
report
source
tables
.gitignore
README.org

How I work - during the project

Create a new folder containing sub-folders:

- `code`: R code
- `data`: processed data
- `source`: original data
- `report`, `figure`, `tables`: output of the analysis

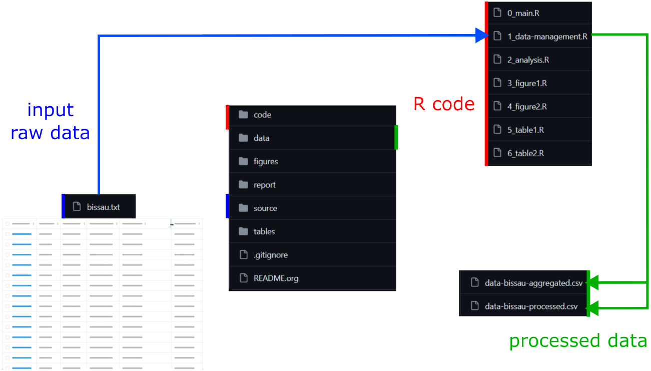


code
data
figures
report
source
tables
.gitignore
README.org

In the `code` folder, I like to have:

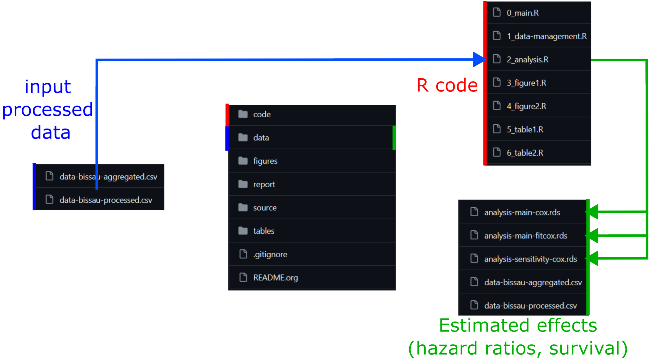
- separate data processing R file(s) + export processed data
- a data R analysis file per research question
  - **with results reported in the article as comments**
  - extra analyses at the end or in a separate file
- `data.frame` objects summarizing results and self sufficient to generate tables and figures.

Structure of the project - data management



- Set the working directory once for all in the main script (`0-main.R`). Other paths should be relative.

Structure of the project - data analysis



input processed data

- data-bissau-aggregated.csv
- data-bissau-processed.csv

- code
- data
- figures
- report
- source
- tables
- .gitignore
- README.org

R code

- 0_main.R
- 1_data-management.R
- 2_analysis.R
- 3_figure1.R
- 4_figure2.R
- 5_table1.R
- 6_table2.R

- analysis-main-cox.rds
- analysis-main-fitcox.rds
- analysis-sensitivity-cox.rds
- data-bissau-aggregated.csv
- data-bissau-processed.csv

Estimated effects (hazard ratios, survival)

- Read processed data, run some analysis, export results.
- Results can then be read and transformed into tables or figures.

How I work - end of the project

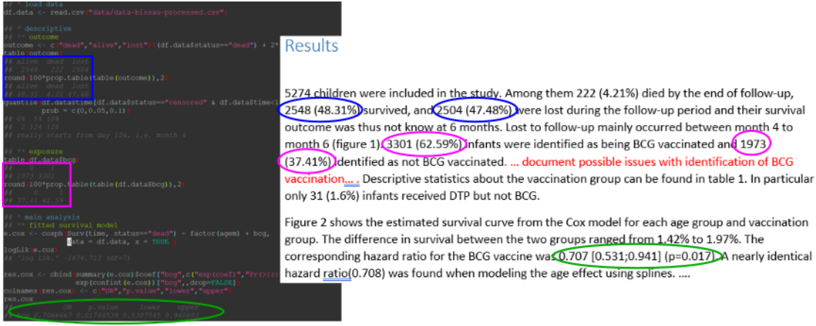
Clean up the project folder

- to reproduce results, tables, and figures of an article **all of those but only those**.
- additional dataset/analyses should removed or put in separate files labeled `EXTRA-...`

This means:

- removing un-necessary data processing/analysis instructions
- creating a separate file for generating each figure and table
- exporting intermediate results used to generate figures & tables (no need to refit complex model to change the caption!)

analysis.R - end of the project



```
## * load data
df.data <- read.csv("data/data-bissau-processed.csv")

## * descriptive
## ** outcome
outcome <- c("dead","alive","lost")[(df.data$status=="dead") + 2*
table(outcome)
## alive  dead  lost
##  2548   222  2504
round(100*prop.table(table(outcome)),2)
## alive  dead  lost
## 48.31  4.21 47.48
quantile(df.data$time[df.data$status=="censored" & df.data$time<1
         prob = c(0,0.05,0.1))
## 0%  5% 10%
##  2 124 128
## really starts from day 124, i.e. month 4

## ** exposure
table(df.data$bcg)
##    0    1
## 1973 3301
round(100*prop.table(table(df.data$bcg)),2)
##     0     1
## 37.41 62.59

## * main analysis
## ** fitted survival model
e.cox <- coxph(Surv(time, status=="dead") ~ factor(agem) + bcg,
               data = df.data, x = TRUE )
logLik(e.cox)
## 'log Lik.' -1876.713 (df=7)

res.cox <- cbind(summary(e.cox)$coef["bcg",c("exp(coef)","Pr(>|z|
                 exp(confint(e.cox))["bcg",,drop=FALSE])
colnames(res.cox) <- c("OR","p.value","lower","upper")
res.cox
##            OR    p.value     lower     upper
## bcg 0.7066667 0.01744539 0.5307545 0.940883
```

Results

5274 children were included in the study. Among them 222 (4.21%) died by the end of follow-up, 2548 (48.31%) survived, and 2504 (47.48%) were lost during the follow-up period and their survival outcome was thus not know at 6 months. Lost to follow-up mainly occurred between month 4 to month 6 (figure 1). 3301 (62.59%) infants were identified as being BCG vaccinated and 1973 (37.41%) identified as not BCG vaccinated. ... document possible issues with identification of BCG vaccination... . Descriptive statistics about the vaccination group can be found in table 1. In particular only 31 (1.6%) infants received DTP but not BCG.

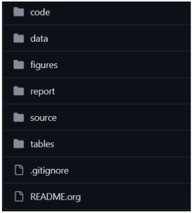
Figure 2 shows the estimated survival curve from the Cox model for each age group and vaccination group. The difference in survival between the two groups ranged from 1.42% to 1.97%. The corresponding hazard ratio for the BCG vaccine was 0.707 [0.531;0.941] (p=0.017). A nearly identical hazard ratio(0.708) was found when modeling the age effect using splines. ....

Should be an obvious connexion between one R file and the result section in the article.

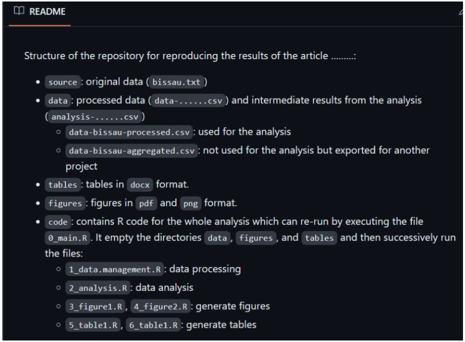
- in more complex projects, one may have: analysis.R for fitting complex models and results.R for presenting results.

Structure of the project - sharing the code

- add `README` file with the specifics about the program used



- code
- data
- figures
- report
- source
- tables
- .gitignore
- README.org



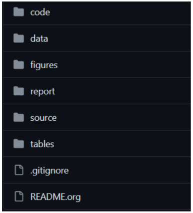
README

Structure of the repository for reproducing the results of the article ....:

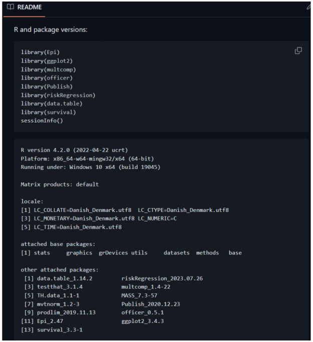
- `source` : original data (`bissau.txt`)
- `data` : processed data (`data-......csv`) and intermediate results from the analysis (`analysis-......csv`)
    - `data-bissau-processed.csv` : used for the analysis
    - `data-bissau-aggregated.csv` : not used for the analysis but exported for another project
- `tables` : tables in `docx` format.
- `figures` : figures in `pdf` and `png` format.
- `code` : contains R code for the whole analysis which can re-run by executing the file `0_main.R` . It empty the directories `data` , `figures` , and `tables` and then successively run the files:
    - `1_data.management.R` : data processing
    - `2_analysis.R` : data analysis
    - `3_figure1.R` , `4_figure2.R` : generate figures
    - `5_table1.R` , `6_table1.R` : generate tables

Structure of the project - sharing the code

- add `README` file with the specifics about the program used



```
code
data
figures
report
source
tables
.gitignore
README.org
```



README

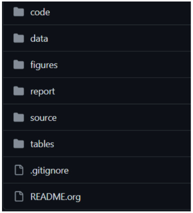
R and package versions:

```
library(Epi)
library(ggplot2)
library(multcomp)
library(officer)
library(Publish)
library(riskRegression)
library(data.table)
library(survival)
sessionInfo()
```

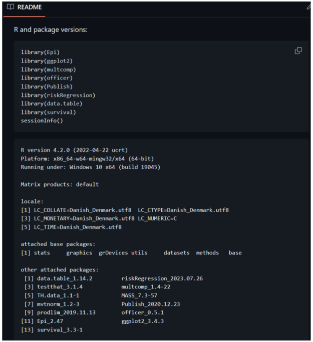
```
R version 4.2.0 (2022-04-22 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

locale:
[1] LC_COLLATE=Danish_Denmark.utf8  LC_CTYPE=Danish_Denmark.utf8
[3] LC_MONETARY=Danish_Denmark.utf8 LC_NUMERIC=C
[5] LC_TIME=Danish_Denmark.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] data.table_1.14.2        riskRegression_2023.07.26
 [3] testthat_3.1.4           multcomp_1.4-22
 [5] TH.data_1.1-1            MASS_7.3-57
 [7] mvtnorm_1.2-3            Publish_2020.12.23
 [9] prodlim_2019.11.13       officer_0.5.1
[11] Epi_2.47                 ggplot2_3.4.3
[13] survival_3.3-1
```

Structure of the project - sharing the code

- add `README` file with the specifics about the program used



```
code
data
figures
report
source
tables
.gitignore
README.org
```



README

R and package versions:

```
library(Epi)
library(ggplot2)
library(multcomp)
library(officer)
library(Publish)
library(riskRegression)
library(data.table)
library(survival)
sessionInfo()
```

```
R version 4.2.0 (2022-04-22 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

locale:
[1] LC_COLLATE=Danish_Denmark.utf8  LC_CTYPE=Danish_Denmark.utf8
[3] LC_MONETARY=Danish_Denmark.utf8 LC_NUMERIC=C
[5] LC_TIME=Danish_Denmark.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] data.table_1.14.2        riskRegression_2023.07.26
 [3] testthat_3.1.4           multcomp_1.4-22
 [5] TH.data_1.1-1            MASS_7.3-57
 [7] mvtnorm_1.2-3            Publish_2020.12.23
 [9] prodlim_2019.11.13       officer_0.5.1
[11] Epi_2.47                 ggplot2_3.4.3
[13] survival_3.3-1
```

The folder should then contain all the necessary information to reproduce the statistical analysis.

⚠ carefully consider whether to include `data` and `source` when sharing the folder

Robustness

(comparable results when varying modeling assumptions)

Sensitivity analyses for the case study?

Main research question

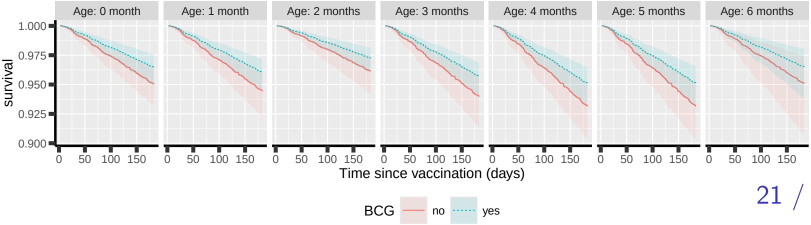
- assessing the effect of BCG on 6-month survival.

Statistical model:

- Cox adjusted for age as categorical variable
  → proportional hazard assumption
  → independent censoring conditional on age and BCG
  → ignore other vaccines

Results: average difference in survival

- -0.0157 [-0.029;-0.003] (p=0.0185)



Possible sensitivity analyses analyses

Changing the covariate set:

1. adding `dtp`, possibly with an interaction with `bcg`.

2. using a spline instead of a categorical age effect.

3. removing `agem` in from the covariates.

Changing the model:

4. using a Kaplan Meier estimator in each vaccination group.

5. using an IPCW logistic regression adjusted for `agem`.

...

Possible sensitivity analyses analyses

Changing the covariate set:

1. adding `dtp`, possibly with an interaction with `bcg`.

2. using a spline instead of a categorical age effect.

3. removing `agem` in from the covariates.

Changing the model:

4. using a Kaplan Meier estimator in each vaccination group.

5. using an IPCW logistic regression adjusted for `agem`.

...

- What are we achieving?
- What is the success/failure criteria?

What are we achieving?

Sensitivity analyses often test how sensitive results are to an assumption:

- start by stating the assumption that is being challenged
  `Proportion hazard assumption`
- then explain how do you challenge it:
  avoid the assumption or make another assumption
  `Kaplan Meier stratified on age and bcg`
  `or IPCW logistic regression`

They can also test how sensitive results are to an arbitrary choice:

- e.g. choice of the age groups, definition of the study population

What are we achieving?

Sensitivity analyses often test how sensitive results are to an assumption:

- start by stating the assumption that is being challenged
  `Proportion hazard assumption`
- then explain how do you challenge it:
  avoid the assumption or make another assumption
  `Kaplan Meier stratified on age and bcg`
  `or IPCW logistic regression`

They can also test how sensitive results are to an arbitrary choice:

- e.g. choice of the age groups, definition of the study population

Sensitivity analyses making unrealistic assumptions, e.g. ignoring known confounders, are in my opinion not relevant.

What is the success/failure criteria?

Comparing p-values between analyses is generally a bad idea

- compare confidence intervals **IF** same underlying estimand

  ```
  0.0156 [0.003;0.029] (p=0.018)
  vs.  0.0166 [0.002;0.031] (p=0.0215)
  ```

Success/failure should be interpreted in the light of how challenging the sensitivity analysis is.

- informally: getting p=0.1 instead of 0.04 when excluding one subject vs. looking at one subgroup.

Beward of non-collapsible estimands (Daniel et al., 2021)

Consider an ideal randomized study:

- no confounding, only age effect on outcome
- no drop-out nor competing risks (death)
- very large sample size

Hazard ratio (HR):

- HR=2.72 in a Cox model with age
- HR=1.89 in a Cox model without age

but nearly identical estimated average survival difference.

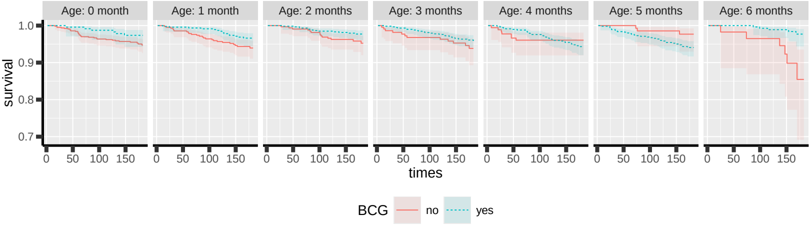
Possible robustness analyses - revisited

Estimand: difference in 6 month survival had all chidren received bcg vs. non received bcg (all else kept equal)

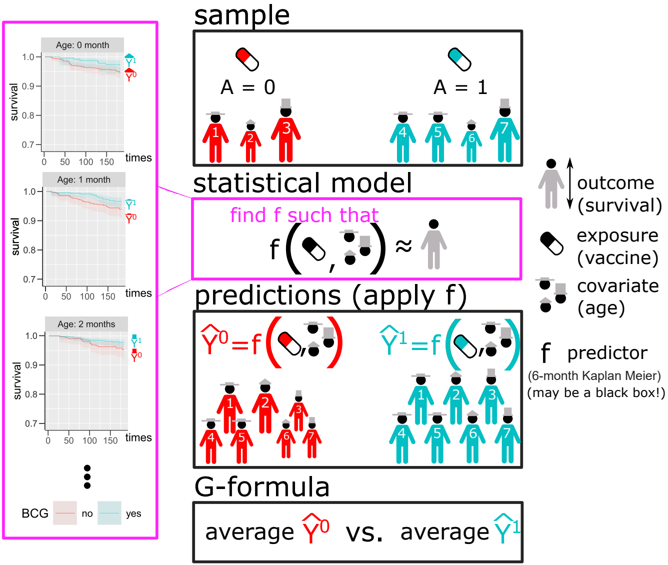
Sensitivity analysis 1: relax the proportional hazard (PH) assumption

- by using a more flexible model: Kaplan Meier estimator stratified on vaccination group and age.

⚠ stratifying only on group would be confusing: relax one assumption (PH) but is exposed to confounding by age.



How to summarize age-specific vaccine effects?



sample

A = 0

A = 1

statistical model

find f such that

$f(\cdot,\cdot) \approx$

predictions (apply f)

$\hat{Y}^0 = f(\cdot,\cdot)$

$\hat{Y}^1 = f(\cdot,\cdot)$

G-formula

average $\hat{Y}^0$ vs. average $\hat{Y}^1$

outcome (survival)

exposure (vaccine)

covariate (age)

f predictor

(6-month Kaplan Meier)

(may be a black box!)

Standardization - by hand

For ease of exposition

- risk = 1-survival
- only consider the first 3 age groups

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	$r_{.,\text{no}} =$	$r_{.,\text{yes}} =$	$n_{.,\text{no}} = 1379$, $n_{.,\text{yes}} = 1303$

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

Standardization - by hand

For ease of exposition

- risk = 1-survival
- only consider the first 3 age groups

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	$r_{.,\text{no}} = 4.37\%$	$r_{.,\text{yes}} = 2.28\%$	$n_{.,\text{no}} = 1379,\ n_{.,\text{yes}} = 1303$

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

Replicablity

(comparable results with another dataset)

- what to replicate

- facilitating replication

Agreement, disagreement, what to replicate?

The vaccine effect in the Guinea-Bissau study:

- was estimated to be 0.0157
- with a standard error (SE) of 0.0067
- with a confidence interval (CI) of $[0.003; 0.029]$
- with a p-value of 0.0185

Consider a randomized study where the vaccine effect:

- was estimated to be -0.0100
- with a standard error (SE) of 0.0344
- with a confidence interval (CI) of $[-0.077; 0.057]$
- with a p-value of 0.771

(same estimand: standardized difference in 6 months survival)

Agreement, disagreement, what to replicate?

The vaccine effect in the Guinea-Bissau study: n=5274

- was estimated to be 0.0157
- with a standard error (SE) of 0.0067
- with a confidence interval (CI) of $[0.003; 0.029]$
- with a p-value of 0.0185

Consider a randomized study where the vaccine effect: n=200

- was estimated to be -0.0100
- with a standard error (SE) of 0.0344
- with a confidence interval (CI) of $[-0.077; 0.057]$
- with a p-value of 0.771

(same estimand: standardized difference in 6 months survival)

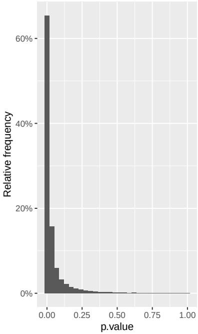
Replicating p-values?

Consider the perfect study:

- 80% power under mean difference of 1
- true mean difference is 1
- no censoring, single age group

P-value distribution in replication studies:
(same sample size)

```
(0,0.001] (0.001,0.005]  (0.005,0.01]
      26.42%        20.31%        10.31%

(0.01,0.05]  (0.05,0.1]     (0.1,1]
     23.68%        7.85%       11.42%
```



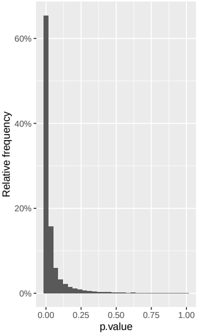
Replicating p-values?

Consider the perfect study:

- 80% power under mean difference of 1
- true mean difference is 1
- no censoring, single age group

P-value distribution in replication studies:
(same sample size)

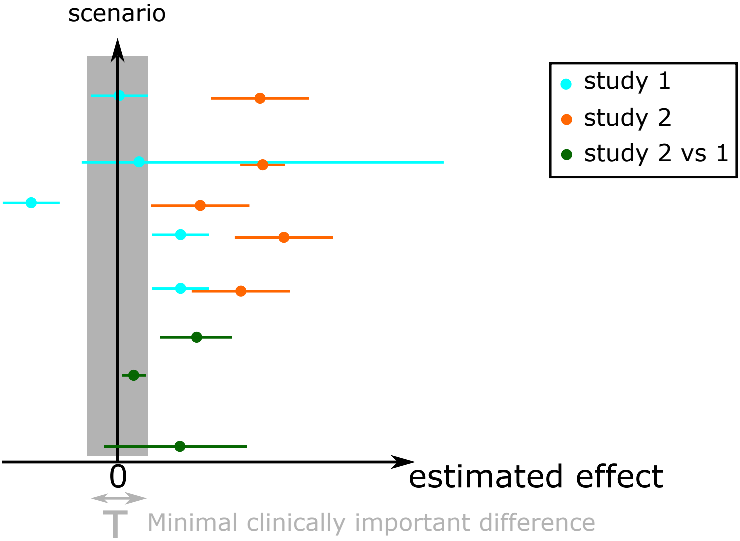
```
(0,0.001] (0.001,0.005]  (0.005,0.01]
      26.42%        20.31%        10.31%

(0.01,0.05]  (0.05,0.1]     (0.1,1]
     23.68%        7.85%      11.42%
```



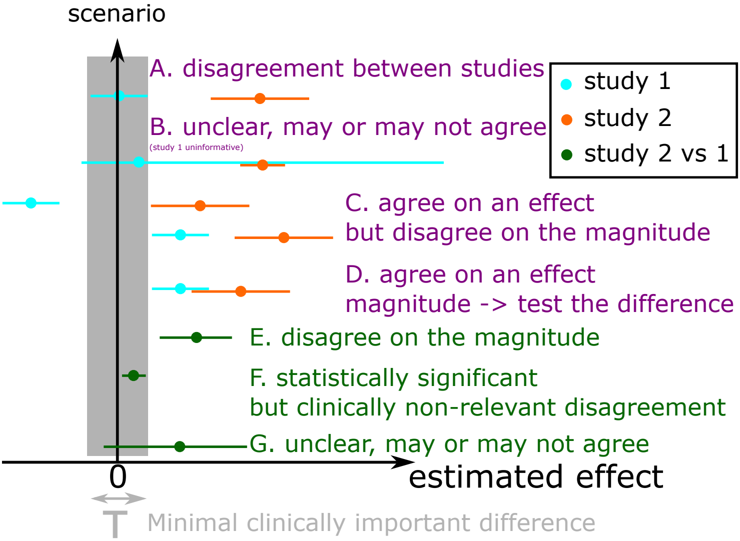
If the observed p-value is 0.0185:

- 62% replication studies lead to a similar p-value (0.001;0.1).
- 80% replication studies have a p-value below 0.05
  (drops to 56% if true mean difference is 0.75)

Look at confidence intervals instead!



Look at confidence intervals instead!


scenario
A. disagreement between studies
B. unclear, may or may not agree
(study 1 uninformative)
C. agree on an effect
but disagree on the magnitude
D. agree on an effect
magnitude -> test the difference
E. disagree on the magnitude
F. statistically significant
but clinically non-relevant disagreement
G. unclear, may or may not agree
0
estimated effect
Minimal clinically important difference
● study 1
● study 2
● study 2 vs 1

- sceario A and B may produce similar p-values

A recent proposal (Spence and Stanley, 2024)

"Unfortunately, many popular and readily accessible methods for ascertaining replicability [. . .] are generally ill suited to the task of comparing results across studies. To address this issue, we present the prediction interval as a statistic that is effective for determining whether a replication study is inconsistent with the original study."

```
library(predictionInterval) ## survival in %
out <- pi.m(M = 1.57, SD = 0.67*sqrt(5274),
            n = 5274, rep.n = 200)
```

```
Original study: M = 1.57, SD = 48.66, N = 5274, 95% CI[0.26, 2.88]
 Replication study: N = 200
 Prediction interval: 95% PI[-5.30,8.44].
 [...]
```

A recent proposal (Spence and Stanley, 2024)

"Unfortunately, many popular and readily accessible methods for ascertaining replicability [...] are generally ill suited to the task of comparing results across studies. To address this issue, we present the prediction interval as a statistic that is effective for determining whether a replication study is inconsistent with the original study."

```
library(predictionInterval) ## survival in %
out <- pi.m(M = 1.57, SD = 0.67*sqrt(5274),
            n = 5274, rep.n = 200)
```

```
Original study: M = 1.57, SD = 48.66, N = 5274, 95% CI[0.26, 2.88]
 Replication study: N = 200
 Prediction interval: 95% PI[-5.30,8.44].
 [...]
```

- depends on the size of the replication study 🥳
- equivalent to a two sample t-test assuming a common variance between the two studies, estimated based on study 1.

Prediction intervals for different sample sizes

```
out <- pi.m(M = 1.57, SD = 0.67*sqrt(5274),
            n = 5274, rep.n = 5274)
```

```
Original study: M = 1.57, SD = 48.66, N = 5274, 95% CI[0.26, 2.88]
 Replication study: N = 5274
 Prediction interval: 95% PI[-0.29,3.43].
 [...]
```

```
out <- pi.m(M = 1.57, SD = 0.67*sqrt(5274),
            n = 5274, rep.n = 2.5*5274)
```

```
Original study: M = 1.57, SD = 48.66, N = 5274, 95% CI[0.26, 2.88]
 Replication study: N = 13185
 Prediction interval: 95% PI[0.02,3.12].
 [...]
```

Facilitating replicability

⚠ Making the analysis reproducible is not enough to make replication possible:

- source code can be very difficult to understand!
- does not age well - software evolve over time
- can be hard or impossible to re-run the code to obtain specific results (e.g. standard error)

This is why we need both:

- description in the method section in plain english
  → Statistical analysis paragraph
- enough results: not only p-values but also estimate and CIs (SE can often be deduced from CIs)
- the source code (with comments) and information about the software for reproducibility (see 'Reproducibility')

Statistical analysis paragraph

Relate the research question(s) to a procedure producing the numbers reported in the result section.

To be exhaustive it should explicit

Statistical analysis paragraph

Relate the research question(s) to a procedure producing the numbers reported in the result section.

To be exhaustive it should explicit

- the **parameter of interest** (or estimand)

Statistical analysis paragraph

Relate the research question(s) to a procedure producing the numbers reported in the result section.

To be exhaustive it should explicit

- the **parameter of interest** (or estimand)
- the **statistical model** (i.e. assumptions) and the dataset used to estimate the model parameters

Statistical analysis paragraph

Relate the research question(s) to a procedure producing the numbers reported in the result section.

To be exhaustive it should explicit

- the **parameter of interest** (or estimand)
- the **statistical model** (i.e. assumptions) and the dataset used to estimate the model parameters
- the **estimation** procedure for the model parameters and the parameter of interest (e.g. Maximum Likelihod Estimation)

Statistical analysis paragraph

Relate the research question(s) to a procedure producing the numbers reported in the result section.

To be exhaustive it should explicit

- the **parameter of interest** (or estimand)
- the **statistical model** (i.e. assumptions) and the dataset used to estimate the model parameters
- the **estimation** procedure for the model parameters and the parameter of interest (e.g. Maximum Likelihod Estimation)
- the **statistical inference** framework:
  - null hypothesis ($\mathcal{H}_0$)
  - type of test (Wald test, likelihood ratio test ...)
  - uncertainty quantification (e.g. non-parametric boostrap asymptotic theory for MLE, ...)
  - adjustment for multiple comparisons

Statistical analysis paragraph (short)

Relate the research question(s) to a procedure producing the numbers reported in the result section.

It should at least contain

- the **parameter of interest**
- the **statistical model** (i.e. assumptions) and the dataset used to estimate the model parameters
- the **statistical inference** framework: adjustment for multiple comparisons

Consider having an online appendix where you provide all the necessary details instead of trying to squeeze a lot of information in half a page.

- space requirement should not be an excuse in 2026!

Discussion

Take home messages

A reproducible analysis is important for yourself and for science

- being organized from the start helps: start with a statistical analysis plan

Be mindful about what is being assessed in sensitivity analyses or when comparing studies:

- same estimand? (⚠ non-collapsibility of hazard/odds rarios)
- ~~p-values~~ → use forest plots showing CIs

Facilitate replicability

- code **AND** statistical paragraph are important!
- report estimates with confidence intervals not only p-values!

Comments or questions?

https://www.goodvibeblog.com/got-mixed-feelings/

Reference I

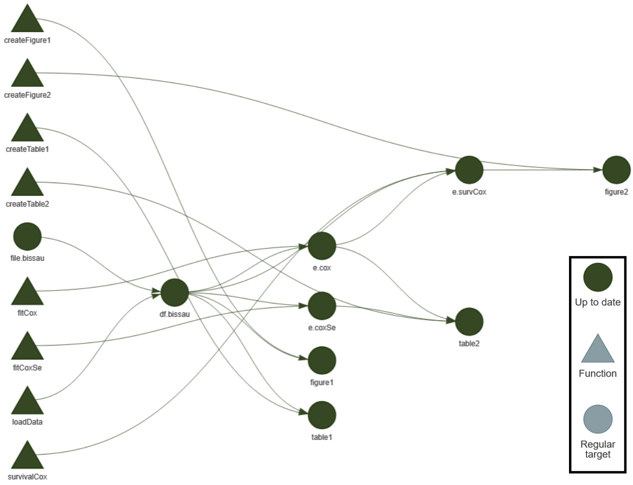
Daniel, R., Zhang, J., and Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3):528–557.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *jama*, 294(2):218–228.

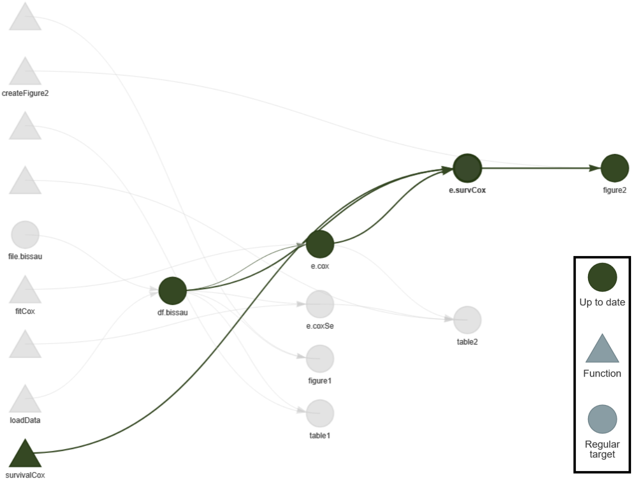
Kristensen, I., Fine, P., Aaby, P., and Jensen, H. (2000). Routine vaccinations and child survival: follow up study in guinea-bissau, west africacommentary: an unexpected finding that needs confirmation or rejection. *Bmj*, 321(7274):1435.

Spence, J. R. and Stanley, D. J. (2024). Tempered expectations: A tutorial for calculating and interpreting prediction intervals in the context of replications. *Advances in Methods and Practices in Psychological Science*, 7(1):25152459231217932.

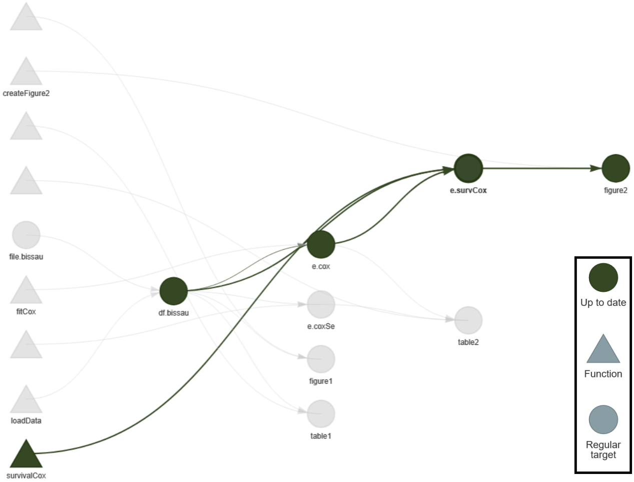
R package targets



package targets



R package targets



- requires to wrap each step into a function with a single output

What do you think about ?

- The outcome was compared between the treatment groups using a Welch's t-test

- A Cox model adjusted on age was used to assess the vaccine effect on survival

What do you think about ?

- The outcome was compared between the treatment groups using a Welch's t-test

- **parameter of interest** : difference in mean outcome

- **statistical model** : group-specific variance,
independent observations

- **estimation** : difference between the empirical means.

- **statistical inference** : $(\mathcal{H}_0)$: equal mean outcome between treatment groups. Wald test with Welch-Satterthwaite approximation for the degree of freedom.

- A Cox model adjusted on age was used to assess the vaccine effect on survival

- unclear, what is the **parameter of interest** ?

A rather comprehensive description

"To estimate the age specific 6 months difference in survival between BCG-vaccinated vs. non vaccinated infant, first a Cox model was used to model the hazard rate of death as a function of time since inclusion in the study (non-parametric baseline hazard), age group (in month as a categorical variable) and BCG vaccination status (yes or no). The BCG vaccination effect was assumed constant over time on the log-hazard scale and identical for all age groups. The p-value relative to the hazard ratio of BCG was used to evaluate the null hypothesis of no vaccine effect at all timepoint in all age groups. The average difference in 6 months survival had all infants been vaccinated vs. none was used to quantify the vaccine effect. The survival was estimated as exponential minus the cumulative hazard with the Breslow estimator of the baseline hazard. "