Teaching responsible statistical analysis practices
&
fostering a positive open science attitude

A learning journey

Martijn Stuiver
Master Evidence Based Practice in Health Care
AmsterdamUMC - University of Amsterdam

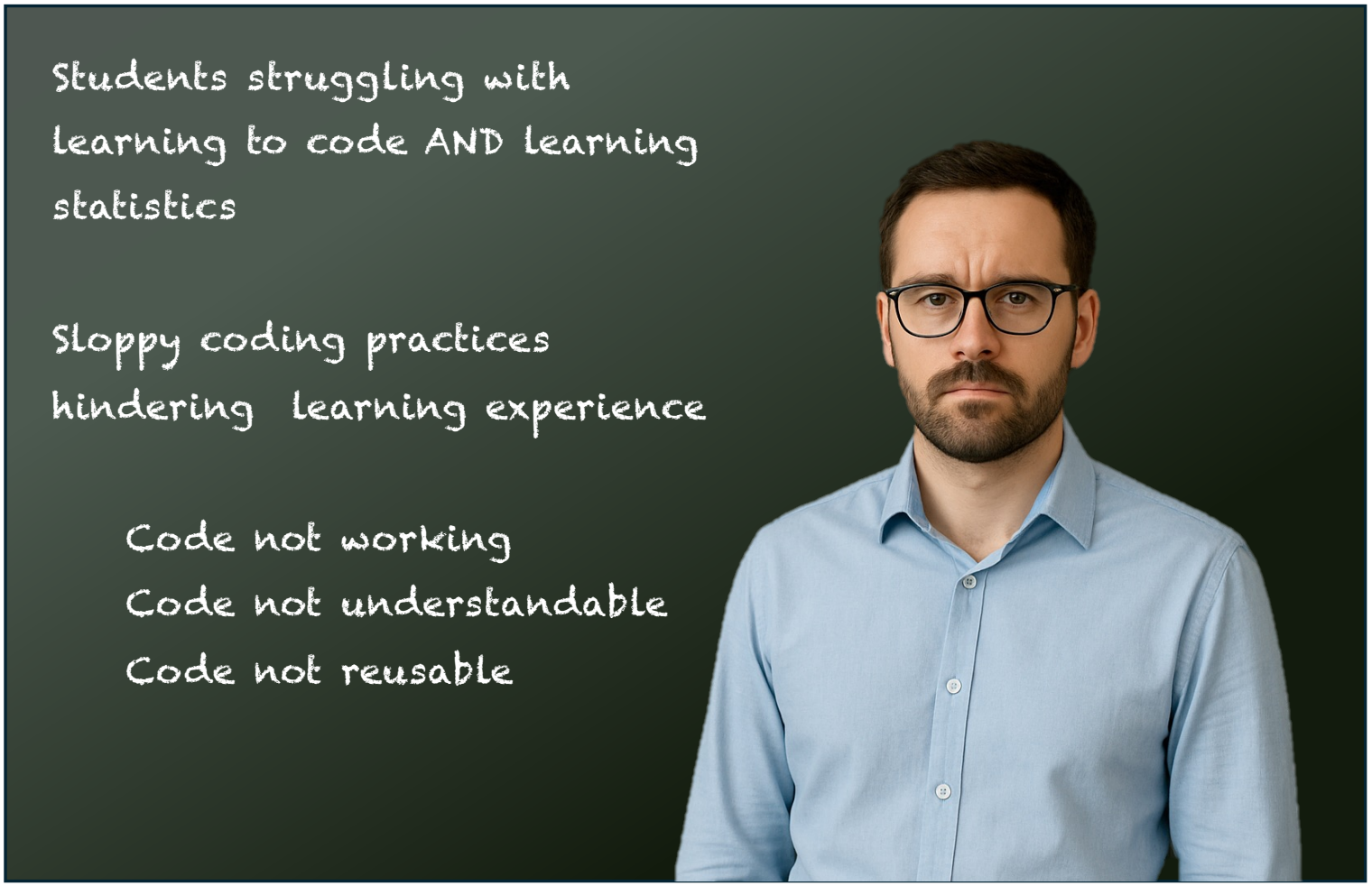
How it started

AI generated image

What we expected

AI generated image

What we got



AI generated image

```r
library(foregn)
libary(foreign)
library(foreign)
mydata <- read.spss(trial.sav, to.data.frame=TRUE)
mydata <- read.spss("trial.sav", to.data.frame=TRUE)
mydata <- read.spss("trial.sav", to.data.frame=TRUE)
mydata <- ("c:/files/course3/trial.sav" to.data.frame=TRUE)
t.test(trial$fvc0~treatm)
t.test(fvc0~treatm, data=trial)
mean(trial$fvc0)
mean(na.omit(trial$fvc0))
sd(na.omit(trial$fvc0))
qqplot(trial$fvc0)
histogram(trial$fvc0)
hist(trial$fvc0)
```

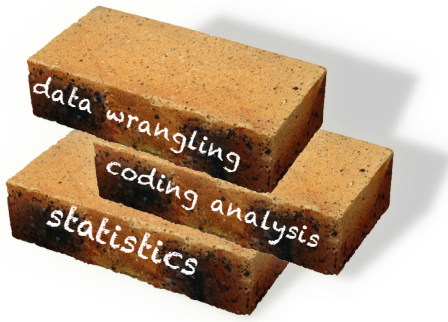
```r
##########################################
# read data to dataframe
##########################################

library(foreign)
mydata <- read.spss("trial.sav", to.data.frame = TRUE)

##########################################
# comparing lung function between groups
##########################################

# first check assumptions

# distribution

mean(trial$fvc0, na.rm=TRUE)
sd(trial$fvco, na.rm=TRUE)
qqplot(trial$fvc0)
hist(trial$fvc0)

# distribution is approximately normal
```

MASTER EBPiHC

Encouraging use of EDC software

- R does not support data *collection*
- Spreadsheet programs come with huge risk of errors
- EDC software aligns with the responsible science paradigm
  - often freely available

- Students required **data wrangling** skills



During thesis research supervision…

Should I check *all* code?

How do I know the code works if I do not have access to the data?

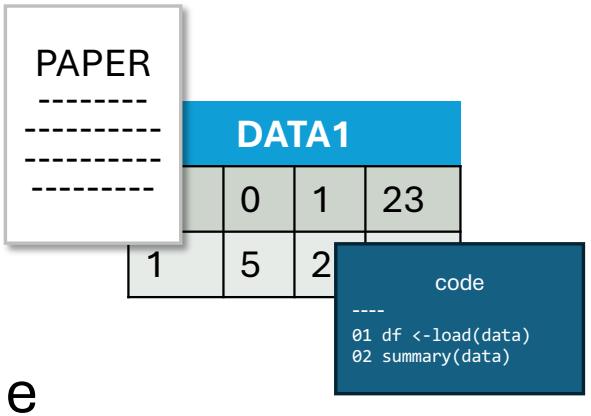
What is this piece of code even supposed to do?

AI generated image

How this relates to open science

Open science is supposed to improve

- Transparency
- Reproducibility
- Reusability



Publish analysis code and (minimal) data alongside research paper → code needs to be comprehensible!

☞ Teaching R presents an opportunity to instill open science attitudes

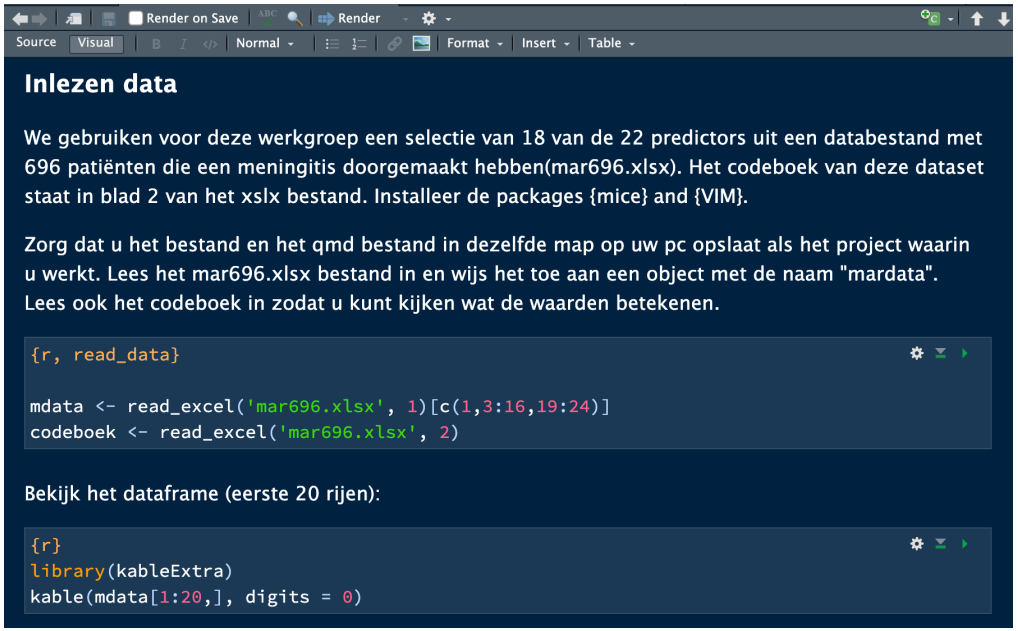
MASTER EBPiHC

RMarkdown

- Write explanation, comments and conclusions
- Embed code
- Embed results
- Knit to good-looking document

- Additional options (i.e. callouts)
- Slightly easier markup option coding
- Supports other programming languages (i.e. python)

MASTER EBPiHC



Inlezen data

We gebruiken voor deze werkgroep een selectie van 18 van de 22 predictors uit een databestand met 696 patiënten die een meningitis doorgemaakt hebben(mar696.xlsx). Het codeboek van deze dataset staat in blad 2 van het xslx bestand. Installeer de packages {mice} and {VIM}.

Zorg dat u het bestand en het qmd bestand in dezelfde map op uw pc opslaat als het project waarin u werkt. Lees het mar696.xlsx bestand in en wijs het toe aan een object met de naam “mardata”. Lees ook het codeboek in zodat u kunt kijken wat de waarden betekenen.

```
mdata <- read_excel('mar696.xlsx', 1)[c(1,3:16,19:24)]
codeboek <- read_excel('mar696.xlsx', 2)
```

Bekijk het dataframe (eerste 20 rijen):

```
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
kable(mdata[1:20,], digits = 0)
```

patnr	cerepals	sex	kldur	epila	antib	gcstot	temp2	bpdiast2	hrate	rash	neck	cranialp	esr	tromb	age	t
10	1	1	1	1	1	11	0	1	120	0	1	0	19	79	47	
2	0	0	1	1	1	15	0	1	135	0	0	1	51	335	19	

Table of contents

Developing a convention I

How to write and organize Quarto files

- YAML frontmatter content
  - Include last date edited
  - Include author
  - Embed resources !
  - Add Table of Content

Use recognizable headings

Write analysis like a “belle histoire”: explain what, why and how in text

Developing a convention II

Include standard code chunks to list:

- R version used
- Packages and version used

- Knit to HTML
  - Looks nice
  - post knitting editing more difficult

```
{r, r_version_and_pkgs_statement}
#| echo: false

cat("Dit document is gegenereerd met R versie:")
getRversion()

attached_pkgs <- sessionInfo()$otherPkgs
pkgs <- data.frame(
  Package = names(attached_pkgs),
  Version = sapply(attached_pkgs, function(x) x$Version),
  row.names = NULL
)

cat("De volgende packages zijn gebruikt:")
pkgs
```

R versie en gebruikte packages

```
Dit document is gegenereerd met R versie:

[1] '4.5.1'

De volgende packages zijn gebruikt:

      Package Version
1      visreg   2.8.0
2         rms   8.1-0
3       Hmisc   5.2-4
4     ggplot2   4.0.0
5        nlme 3.1-168
```

MASTER EBPiHC

Dealing with LLMs

LLMs *will* produce code

Often works

Regularly does something other than students think

Atypical, inconsistent or inefficient code

Problem solving changes from learning to code problems to learning to prompt LLMs

How to foster a critical attitude?

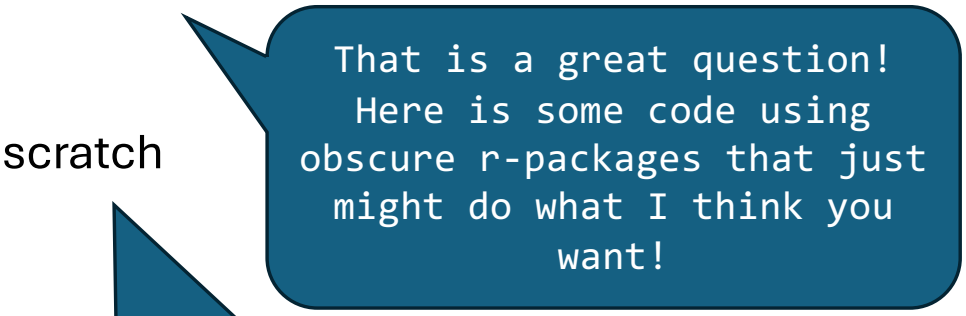
AI generated image

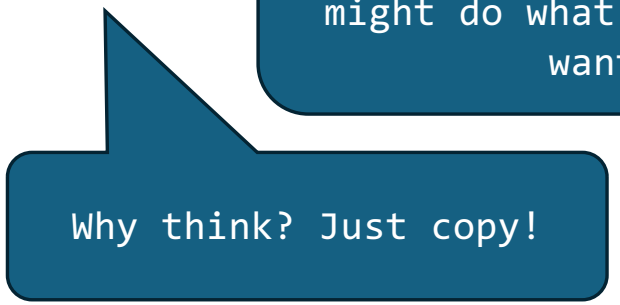
Critical use of LLM generated code

Demonstrate risks and problems associated with AI-generated code

Explain when and how it can be useful

Actively discourage use for generating code from scratch





Let students experiment

- Can they explain the AI generated code?
- Can they demonstrate the code does what it is supposed to do?

Fostering accountability

Results and conclusions discussed with thesis supervisor, with knitted HTML as guiding document

- Student needs to explain approach used
- Teacher can identify unconventional coding and ascertain student understanding

Rewards producing understandable and functioning code

AI generated image

Open access Communication

BMJ Open

Improving reproducibility of data analysis and code in medical research: 5 recommendations to get started

Anna Maria Streiber ,[1,2] Sanne J W Hoepel,[2] Elisabet Blok,[3] Frank J A van Rooij,[2] Julia Neitzel,[1,2] Jeremy Labrecque,[2] M Kamram Ikram,[2,4] Daniel Bos[1,2,5,6]

Reproducibility "Make sure that someone else can reproduce your results"		
Transparency		
Description	**Notes**	☑
Is version control used? *Tip: Use Git or specify the software and package versions.*		☐
Is the input of the code clearly defined? *Tip: Explain what dataset incl. version you use. Ideally this is raw data.*		☐
Is there a ReadMe? *Tip: Explain the structure of your folders, coding of variables, etc.*		☐
Are data cleaning steps reported in the code?		☐
Are sample selection steps (in- and exclusion) reported in the code?		☐

Rewarding transparancy

- R code checklist as part of the thesis judication
- Inspired by Streiber & Hoepel et al.
- Slightly modified to better fit the thesis process

MASTER EBPiHC

Summary

- The transition to R has been challenging but rewarding
- Implementation of Rstudio and Quarto eventually improved (most) students' workflow as intended
- Fostering an open science mindset can be a natural part of statistics courses
- Using "open science workflows" counteracts negative side- effects of the (inevitable) use of LLMs by students
- Teach open science principles by doing instead of preaching

THANK YOU