

# Open science in registry research

Øystein Ariansen Haaland  
Copenhagen, January 2020

# Outline

- Data availability
- Reproducible results
- Sharing of syntax
- Advanced: Three levels of open science

# Data availability

“NIH will [...] require NIH-funded researchers to make the data underlying [...] scientific research publications freely available [...].”

“Investigators are expected to share with other researchers [...] the primary data [...] created or gathered in the course of work under NSF grants.”

# Data availability

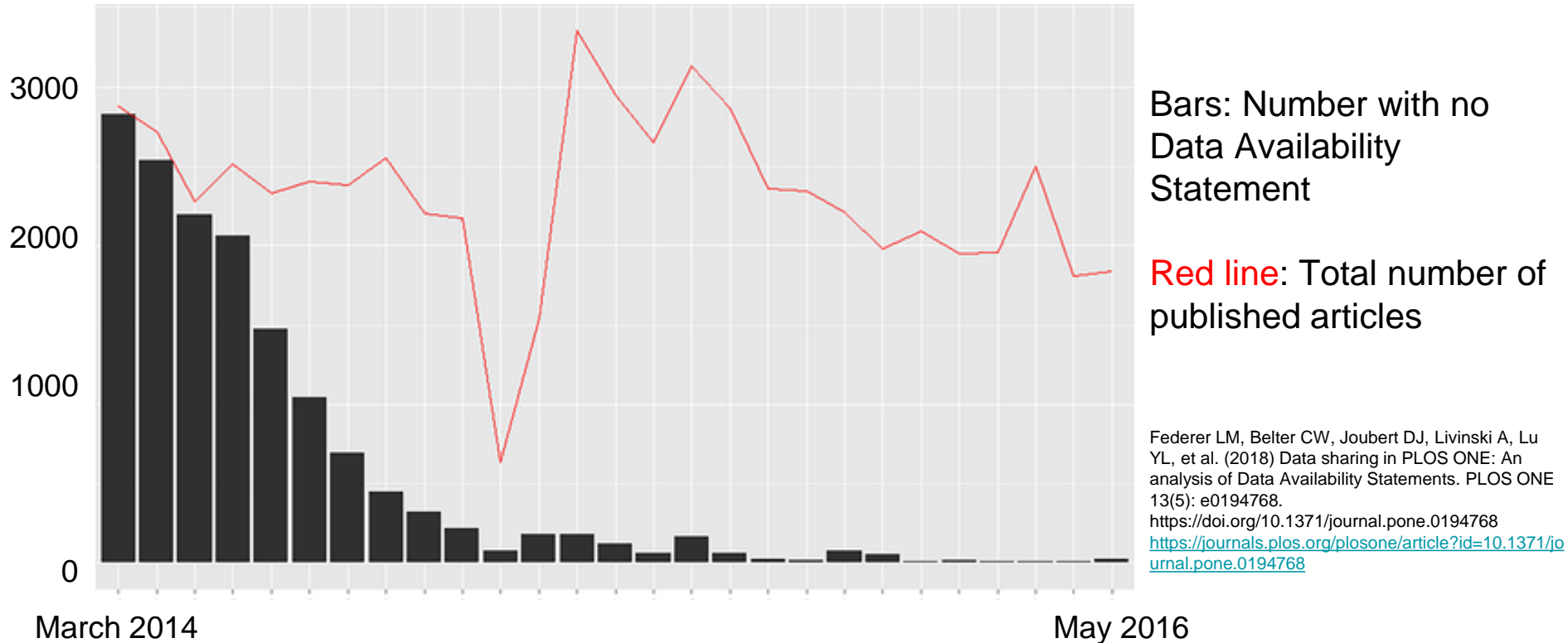
“PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction at the time of publication.”

Elsevier: “Research data should be made available free of charge to all researchers wherever possible and with minimal reuse restrictions.”

Science: “After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of a *Science* Journal.”

Nature: “Supporting data must be made available to editors and peer reviewers at the time of submission for the purposes of evaluating the manuscript.”

# Articles missing Data Availability Statement in PLOS journals



March 2014

May 2016

# Data availability

- In registry research, data sharing is often unethical AND illegal!

# Data availability

- In registry research, data sharing is often unethical AND illegal!

How would you get around this? Take two minutes and talk with your neighbor.

# Data availability

- In registry research, data sharing is often unethical AND illegal!

## RESEARCH ARTICLE

# $\beta$ 2-Adrenoreceptor is a regulator of the $\alpha$ -synuclein gene driving risk of Parkinson's disease

Shuchi Mittal<sup>1,2,3</sup>, Kjetil Bjørnevik<sup>4,5</sup>, Doo Soon Im<sup>6</sup>, Adrian Flierl<sup>7</sup>, Xianjun Dong<sup>1,2,3</sup>, Joseph J. Locascio<sup>1,8</sup>, Kristine M. Ab...

+ See all authors and affiliations

Science 01 Sep 2017:

Vol. 357, Issue 6354, pp. 891-898

DOI: 10.1126/science.aaf3934





# Data availability

- In registry research, data sharing is often unethical AND illegal!

## RESEARCH ARTICLE

### $\beta$ 2-Adrenoreceptor is a regulator of the $\alpha$ -synuclein gene driving risk of Parkinson's disease

Shuchi Mittal<sup>1,2,3</sup>, Kjetil Bjørnevik<sup>4,5</sup>, Doo Soon Im<sup>6</sup>, Adrian Flierl<sup>7</sup>, Xianjun Dong<sup>1,2,3</sup>, Joseph J. Locascio<sup>1,8</sup>, Kristine M. Ab...


+ See all authors and affiliations

Science 01 Sep 2017:  
Vol. 357, Issue 6354, pp. 891-898  
DOI: 10.1126/science.aaf3934

- “NorPD data are accessible by application at <http://norpd.no>.”

# Data availability


## Socio-Economic Status and Reproduction among Adults Born with an Oral Cleft: A Population-Based Cohort Study in Norway

Erik Berg , Åse Sivertsen, Anja Maria Steinsland Ariansen, Charles Filip, Halvard A. Vindenes, Kristin B. Feragen, Dag Moster, Rolv Terje Lie, Øystein A. Haaland

Published: September 15, 2016 • <https://doi.org/10.1371/journal.pone.0162196>

# Data availability

## Socio-Economic Status and Reproduction among Adults Born with an Oral Cleft: A Population-Based Cohort Study in Norway

Erik Berg , Åse Sivertsen, Anja Maria Steinsland Ariansen, Charles Filip, Halvard A. Vindenes, Kristin B. Feragen, Dag Moster, Rolv Terje Lie, Øystein A. Haaland

Published: September 15, 2016 • <https://doi.org/10.1371/journal.pone.0162196>

**“Data Availability:** We are not allowed by Norwegian law to make the data available, as the datasets used in the study contain sensitive patient information. The regional ethical committee for medical and health research ethics does not allow for public deposition of the data. Readers can apply for access and permission to analyze data from each of the registries involved (The Medical Birth Registry: <http://www.fhi.no/artikler/?id=94819>; FD-Trygd <http://www.ssb.no/omssb/tjenester-og-verktoy/data-til-forskning/fd-trygd>).”

# Data availability

“Anonymize” data to avoid reidentification.

# Data availability

## k-anonymization

- Age: 0,1,2,...,100 → 0-4, 5-9, ..., 74-79, 80+  
Length of stay at ICU: hours → whole days  
Date of admission: 21 January 2017 → January 2017  
Cause of death: “Spacecraft accident injuring occupant” (ICD10 V95.4) → “Transport accident”
- Make cross-table of all variables  
No cell should contain  $<k$  individuals  
 $k=5?$
- ASK FOR PERMISSION TO SHARE!

# Data availability

## Fuzzy the data

- Add random noise (age, birth weight, date of birth)
  - May yield bias
  - Probably affects p-values and confidence intervals
  - Risk of “impossible” values
  - Difficult with binary variables (dead vs. alive, male vs. female)
  
- ASK FOR PERMISSION TO SHARE!

# Data availability

Paper from 2017 discussing anonymization

Research Article

Cancer  
Epidemiology,  
Biomarkers  
& Prevention

## Protecting Privacy in Large Datasets—First We Assess the Risk; Then We Fuzzy the Data

Giske Ursin<sup>1,2,3</sup>, Sagar Sen<sup>1,4</sup>, Jean-Marie Mottu<sup>5</sup>, and Mari Nygård<sup>1</sup>



# Reproducible results

Data handling and analyses.

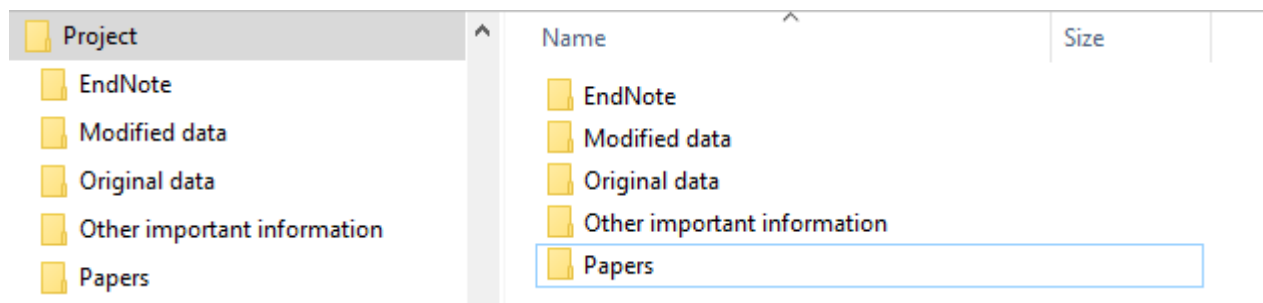
What are important aspects to consider?

Take two minutes and discuss with your neighbor.



# Reproducible results

Organize your data!

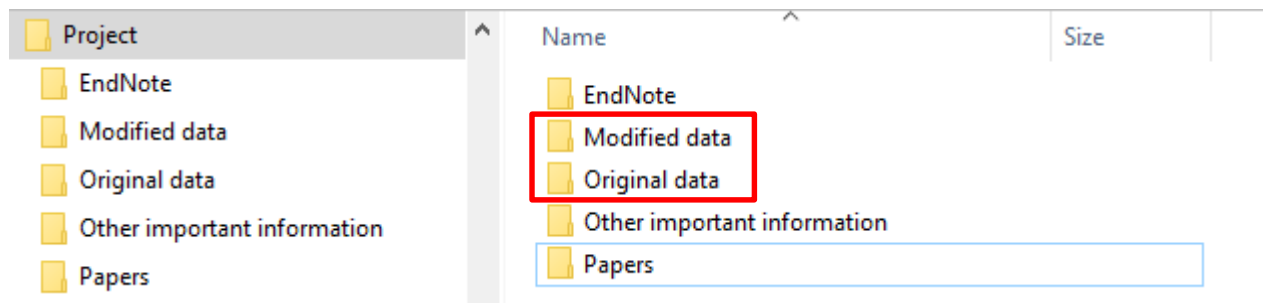


The image shows a file explorer window with a left-hand navigation pane and a main content area. The left pane shows a tree view with a 'Project' folder selected, containing subfolders: 'EndNote', 'Modified data', 'Original data', 'Other important information', and 'Papers'. The main content area displays a table with columns 'Name' and 'Size'. The 'Papers' folder is selected and highlighted with a blue border.

Name	Size
EndNote	
Modified data	
Original data	
Other important information	
Papers	

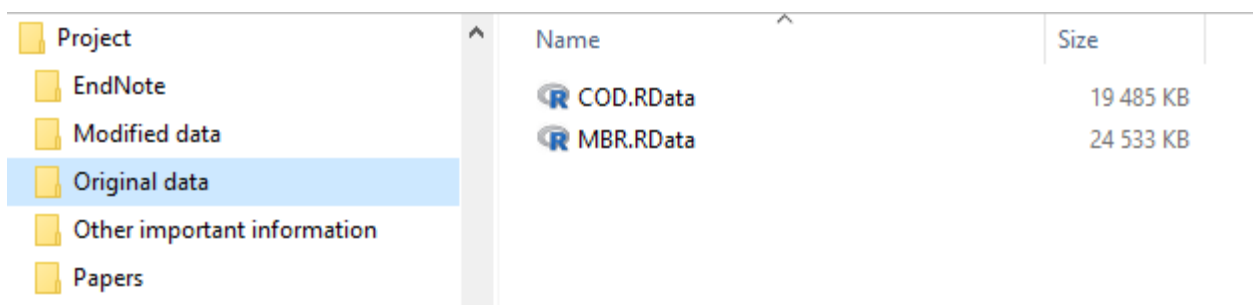
# Reproducible results

Organize your data!



# Reproducible results

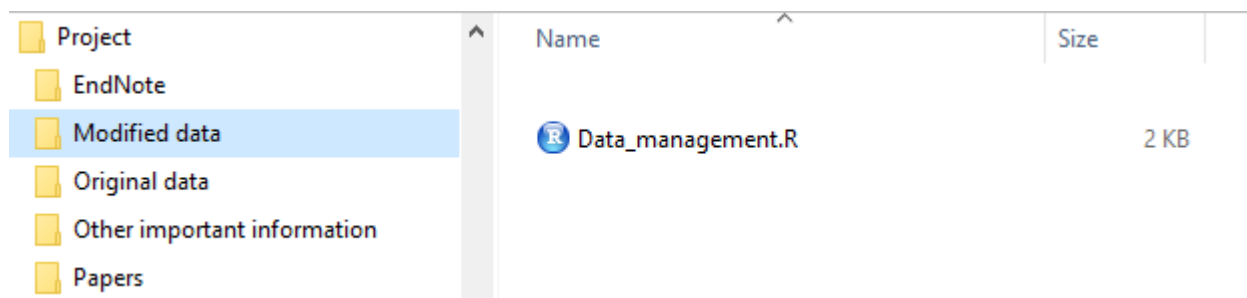
Keep original data separated from your working directory




Name	Size
COD.RData	19 485 KB
MBR.RData	24 533 KB

# Reproducible results

Keep code files and modified data in one folder



	Name	Size
Project		
EndNote		
<b>Modified data</b>	 Data_management.R	2 KB
Original data		
Other important information		
Papers		

# Reproducible results

Make sure that your code always produces the same results

```
Data_management.R x
Source on Save
1 ## Set working directory
2 setwd("C:/Project/Modified data")
3 ## clears workspace
4 rm(list = ls(all.names = T))
5 ## Load data
6 load("../Original data/MBR.RData")
7 load("../Original data/COD.RData")
8 ls()
9
```

# Reproducible results

Make sure that your code always produces the same results

```
Data_management.R x
Source on Save
1 ## Set working directory
2 setwd("C:/Project/Modified data")
3 ## clears workspace
4 rm(list = ls(all.names = T))
5 ## Load data
6 load("../Original data/MBR.RData")
7 load("../Original data/COD.RData")
8 ls()
9
```

Set correct working directory.  
NOT «C:/Project/Original data»

# Reproducible results

Make sure that your code always produces the same results

```
Data_management.R x
Source on Save
1 ## Set working directory
2 setwd("C:/Project/Modified data")
3 ## clears workspace
4 rm(list = ls(all.names = T))
5 ## Load data
6 load("../Original data/MBR.RData")
7 load("../Original data/COD.RData")
8 ls()
9
```

Clear current workspace.  
Avoids that unknown objects are already loaded.

# Reproducible results

Make sure that your code always produces the same results

```
Data_management.R x
Source on Save
1 ## Set working directory
2 setwd("C:/Project/Modified data")
3 ## clears workspace
4 rm(list = ls(all.names = T))
5 ## Load data
6 load("../Original data/MBR.RData")
7 load("../Original data/COD.RData")
8 ls()
9
```

Load original files from source.



# Reproducible results

Make sure that your code always produces the same results

```
Data_management.R x
Source on Save
1 ## Set working directory
2 setwd("C:/Project/Modified data")
3 ## clears workspace
4 rm(list = ls(all.names = T))
5 ## Load data
6 load("../Original data/MBR.RData")
7 load("../Original data/COD.RData")
8 ls()
9
```



# Reproducible results

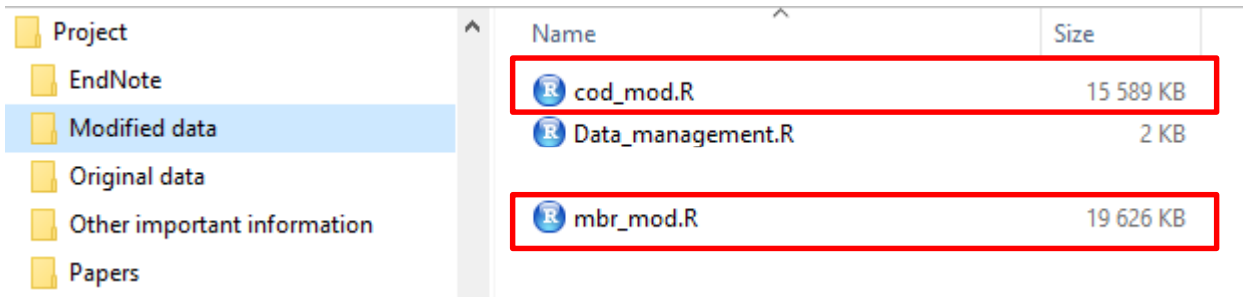
Make sure that your code always produces the same results

```
## Modify mbr and cod
##
## ...
##
## Save modified versions to CURRENT directory
save(mbr_mod,file = "mbr_mod.R")
save(cod_mod,file = "cod_mod.R")
```




Rename modified data sets.  
Do NOT save to «C:/Project/Original data»

# Reproducible results

Make sure that your code always produces the same results



A screenshot of a file explorer interface. On the left, a sidebar shows a tree view of folders: 'Project', 'EndNote', 'Modified data' (highlighted), 'Original data', 'Other important information', and 'Papers'. The main pane displays a table of files with columns 'Name' and 'Size'. Three rows are highlighted with red boxes: 'cod\_mod.R' (15 589 KB), 'Data\_management.R' (2 KB), and 'mbr\_mod.R' (19 626 KB). Each row starts with a blue R logo icon.

Name	Size
 cod_mod.R	15 589 KB
 Data_management.R	2 KB
 mbr_mod.R	19 626 KB

# Reproducible results

Make sure that your code always produces the same results

The image displays a file explorer interface on the left and two screenshots of the Super Mario Bros. game on the right. The file explorer shows a project structure with folders: Project, EndNote, Modified data (highlighted), Original data, Other important information, and Papers. A list of files is shown, with three files highlighted by red boxes: `cod_mod.R`, `Data_management.R`, and `mbr_mod.R`. The top screenshot of the game shows Mario at the top of a green pipe, with a score of 003950, 0 coins, and a time of 300. The bottom screenshot shows Mario in a dark underground level, with a score of 138550, 67 coins, and a time of 168. The NordForsk logo is visible in the bottom right corner.

# Reproducible results

Make sure that your code always produces the same results

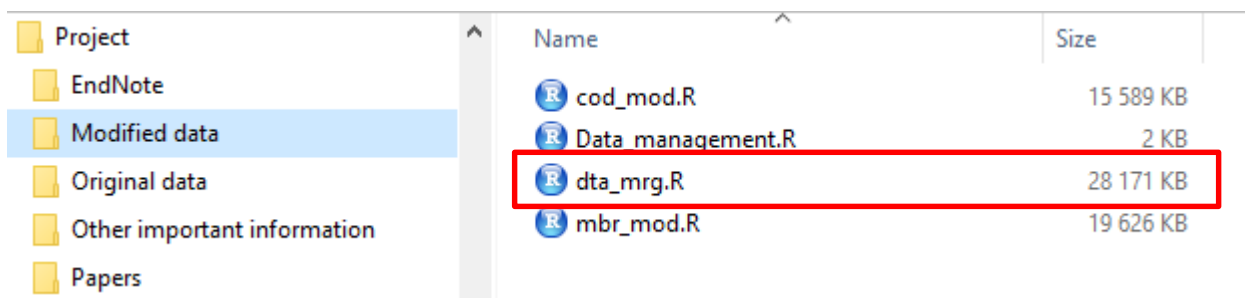
```
## Merge mbr_mod and cod_mod into dta_mrg  
##  
## ...  
##  
## Save merged data to CURRENT directory  
save(dta_mrg, file = "dta_mrg.R")
```





Merge modified data sets  
Modify further  
Again, do NOT save to «C:/Project/Original data»



# Reproducible results

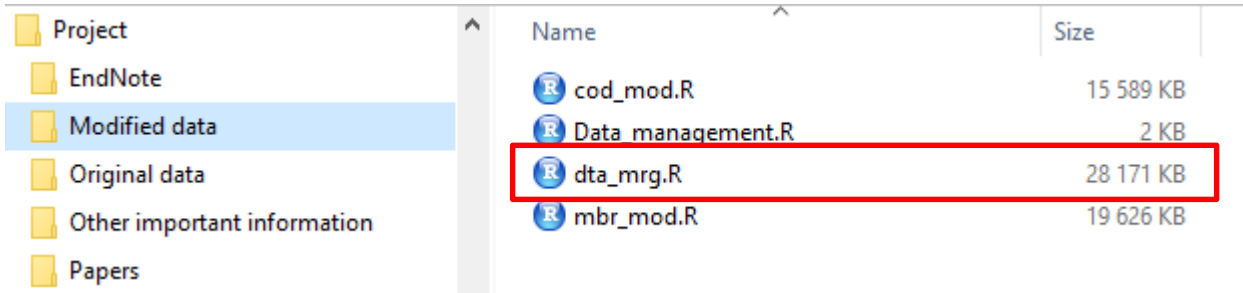
Make sure that your code always produces the same results



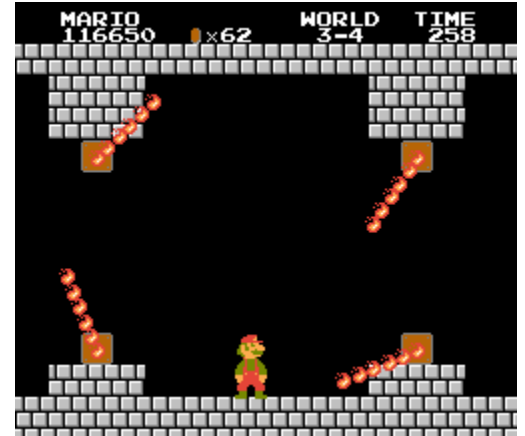
Name	Size
 cod_mod.R	15 589 KB
 Data_management.R	2 KB
 dta_mrg.R	28 171 KB
 mbr_mod.R	19 626 KB

# Reproducible results

Make sure that your code always produces the same results




Name	Size
cod_mod.R	15 589 KB
Data_management.R	2 KB
dta_mrg.R	28 171 KB
mbr_mod.R	19 626 KB



# Reproducible results

## Preparing analyses

```
## Perform analyses
## Start new session by emptying old workspace and load correct data
setwd("C:/Project/Modified data")
rm(list = ls(all.names = T))
## Load data
load("dta_mrg.R")
ls()
##
## ...
## analyses
## ...
```

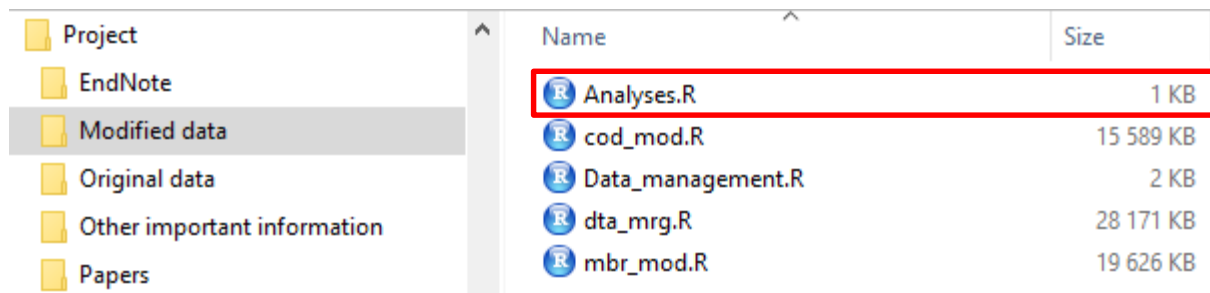


Create new file  
Set working directory  
Clear workspace  
Load correct data set








# Reproducible results

## Preparing analyses



A screenshot of a file explorer window showing a project directory. The left pane displays a tree view with folders: Project, EndNote, Modified data (selected), Original data, Other important information, and Papers. The right pane shows a table of files with columns 'Name' and 'Size'. The file 'Analyses.R' is highlighted with a red box.

Name	Size
 Analyses.R	1 KB
 cod_mod.R	15 589 KB
 Data_management.R	2 KB
 dta_mrg.R	28 171 KB
 mbr_mod.R	19 626 KB

# Reproducible results

## Reporting results

What are good habits to have when reporting results?

Take two minutes to talk with your neighbor.

# Reproducible results

## Reporting results

### Auto-generate tables

- R: `rmarkdown`, `write.table()`, `write.csv()`, `write.xlsx()`
- Stata: `tab2xl`, `tab2docx`, `putdocx`, `dyndoc`

### Write figures to pdf/jpg/tiff/... from script

- R: `pdf()`, `jpeg()`, `tiff()`, ...
- Stata: `putpdf`

### Comment in script why you conduct analyses

- “Regression to test ...”
- “Mean of all ...”

# Sharing of syntax

## Post syntax online

- Repository
- Supplementary information
- Home page

## “Clean” version

- Code that produces the results in the paper
- Good comments (to be understood by yourself in two years)

# Three levels of open science

## Level 0 – minimal

A project website

- Maintained, updated

Deposition of papers

- Accepted drafts
- Publicly available

Deposition of datasets

- Publicly available

Use open-access journals

<http://alanwinfield.blogspot.com/2014/11/open-science-preaching-what-i-practice.html>

# Three levels of open science

Level 1 = Level 0 +:

Regular project blogs

- Respond to feedback

Post relevant videos online

- Explanation and commentary

Engage in social media

- Twitter, Facebook, Instagram, etc.

<http://alanwinfield.blogspot.com/2014/11/open-science-preaching-what-i-practice.html>

# Three levels of open science

Level 2 = Level 1 +:

Daily notebooks written online

- Accessible in real-time

Upload working datasets

- Explanation and commentary

Publicly accessible wiki for project dialogue

<http://alanwinfield.blogspot.com/2014/11/open-science-preaching-what-i-practice.html>

That's all for now