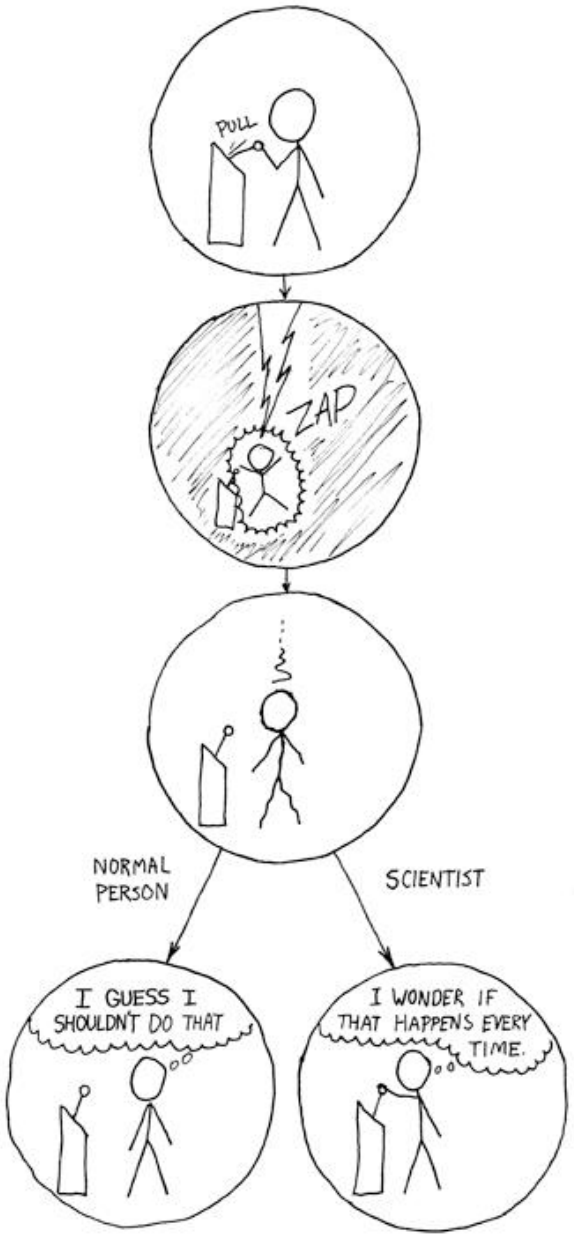
Department of Systems Neuroscience



Methods Meeting

Open and Reproducible Science

Mana Ehlers

March 16, 2021

Universitätsklinikum Hamburg-Eppendorf

Why open science?

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there

---

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡

14 Nov 2013

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]

[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Psychological Science
22(11) 1359–1366

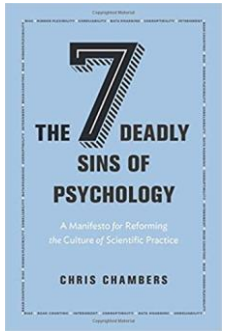
DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hy-



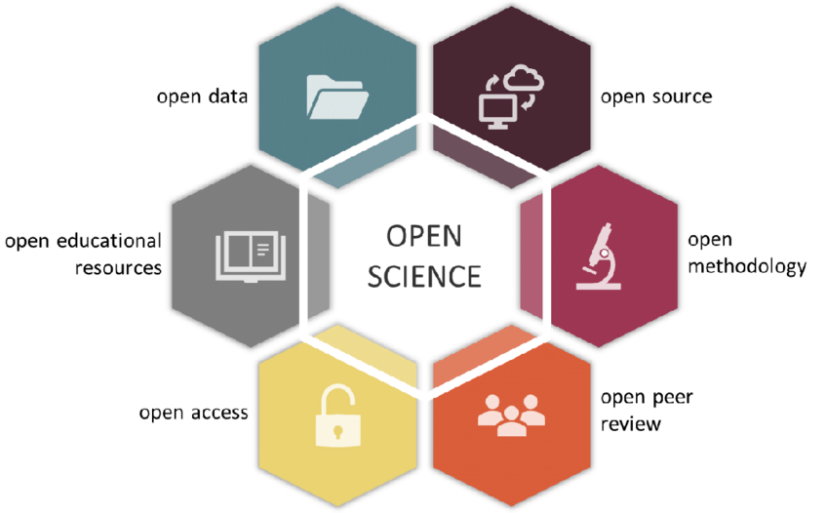
Editorial

Registered Reports

A Method to Increase the Credibility of Published Results

Brian A. Nosek[1] and Daniël Lakens[2]

[1]University of Virginia and Center for Open Science, Charlottesville, VA, USA,
[2]Eindhoven University of Technology, The Netherlands



The preregistration revolution

Brian A. Nosek[a,b,1], Charles R. Ebersole[b], Alexander C. DeHaven[a], and David T. Mellor[a]

[a]Center for Open Science, Charlottesville, VA 22903; and [b]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved August 28, 2017 (received for review June 15, 2017)

Increasing Transparency Through a Multiverse Analysis

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, more…

First Published September 29, 2016 | Research Article | Find in PubMed
https://doi.org/10.1177/1745691616658637

Article information

Abstract

Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

Psychol Sci. 2017 Nov;28(11):1547-1562. doi: 10.1177/0956797617723724. Epub 2017 Sep 13.

Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty

Samantha F Anderson [1], Ken Kelley [1], Scott E Maxwell [1]

Review | Nat Neurosci. 2017 Feb 23;20(3):299-303. doi: 10.1038/nn.4500.

Best practices in data analysis and sharing in neuroimaging using MRI

Thomas E Nichols [1], Samir Das [2 3], Simon B Eickhoff [4 5], Alan C Evans [2 3], Tristan Glatard [2 6], Michael Hanke [7 8], Nikolaus Kriegeskorte [9], Michael P Milham [10 11], Russell A Poldrack [12], Jean-Baptiste Poline [13], Erika Proal [14], Bertrand Thirion [15], David C Van Essen [16], Tonya White [17], B T Thomas Yeo [18]

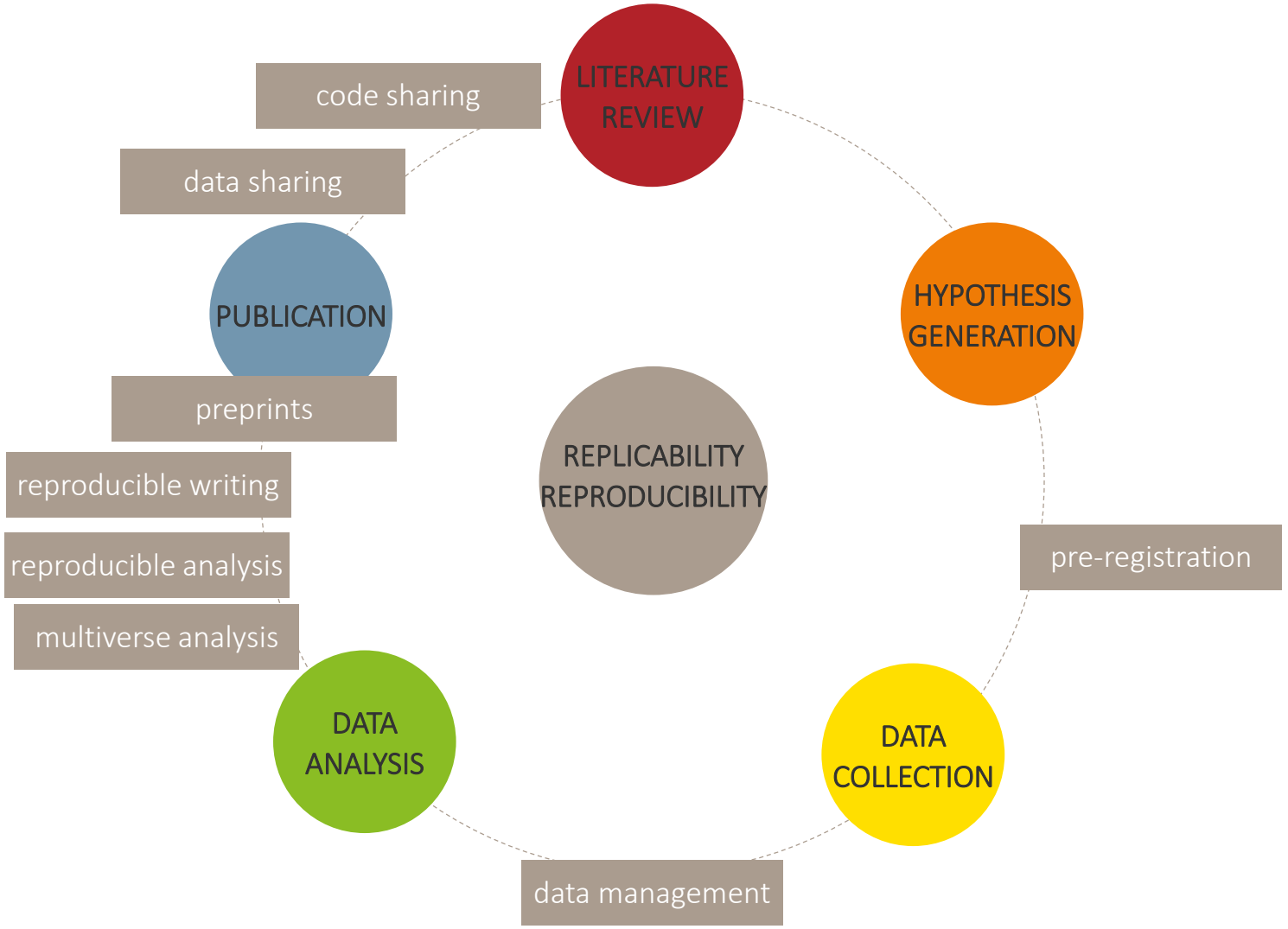
Open Access | Published: 10 January 2017

A manifesto for reproducible science

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

The research process



Pre-registration: Why and What?

Why?

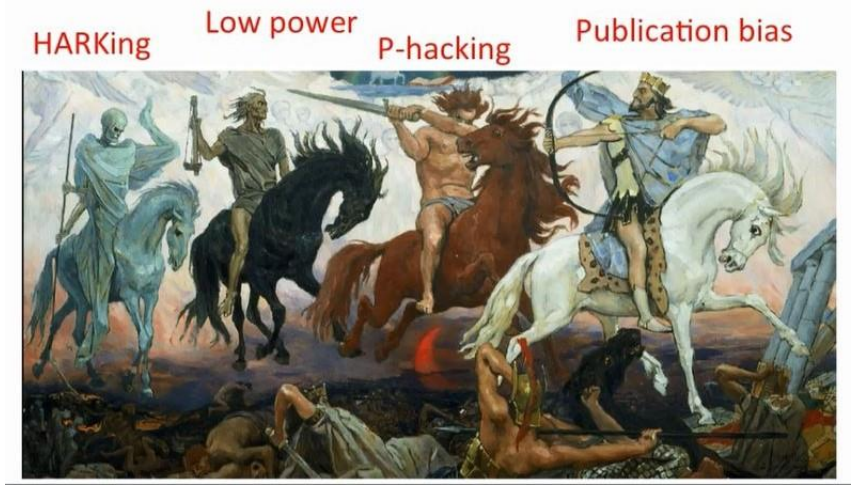
- Counteracts (publication) bias, low power, p-hacking/researchers degrees of freedom, HARKing
- Distinction between exploratory and confirmatory hypotheses
- Promotes transparent research
- Encourages more thorough planning

What?

- time-stamped research plan
- submitted (before data collection) to public registry

Content (minimum)

- Research question and hypotheses
- Dependent and indepenent variables
- Sample size (justification)
- Statistical model specification



The four horsemen of reproducibility – Dorothy Bishop

Pre-registration: How and Where?

1) Open-ended registration: You make the rules! Provide summary of project.

2) Aspredicted.org: 8 questions about project (state of data collection, hypothesis, variables, sample size, analyses, others)

AS PREDICTED

3) OSF pre-registration template: covers 25 subject areas (basics plus room for introduction, design plan, sampling plan, data collection details, exclusions etc.)

OSF PREREGISTRATION

4) Registered report

The process – Hands on

Study Information

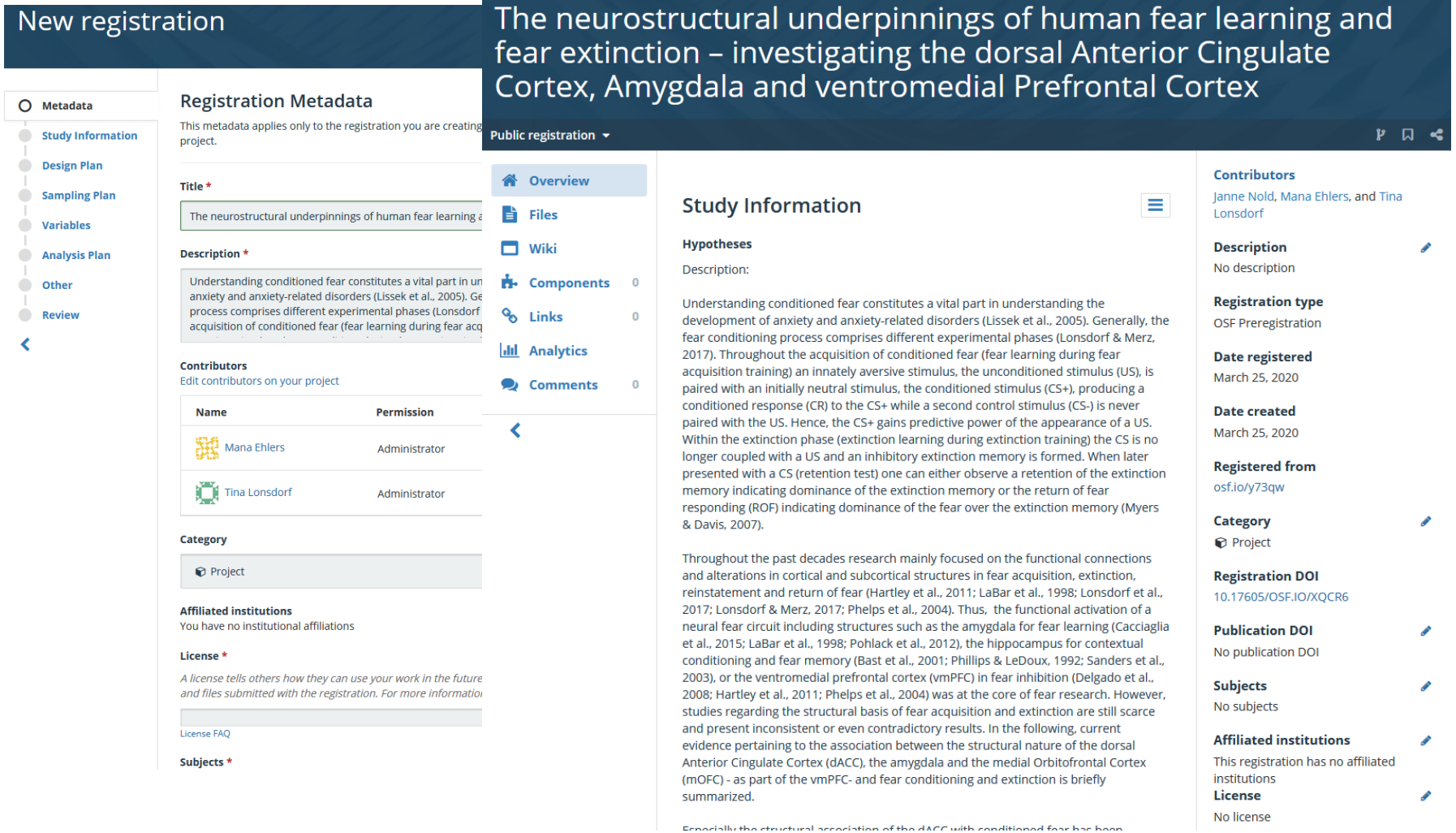
1. Title (required)
   1.1. Provide the working title of your study. It may be the same title that you submit for publication of your final manuscript, but it is not a requirement.
   1.2. **Example**: Effect of sugar on brownie tastiness.
   1.3. **More info**: The title should be a specific and informative description of a project. Vague titles such as 'Fruit fly preregistration plan' are not appropriate.

2. Authors (required)

3. Description (optional)
   3.1. Please give a brief description of your study, including some background, the purpose of the study, or broad research questions.
   3.2. **Example**: Though there is strong evidence to suggest that sugar affects taste preferences, the effect has never been demonstrated in brownies. Therefore, we will measure taste preference for four different levels of sugar concentration in a standard brownie recipe to determine if the effect exists in this pastry.

   3.3. **More info**: The description should be no longer than the length of an abstract. It can give some context for the proposed study, but great detail is not needed here for your preregistration.

4. Hypotheses (required)
   4.1. List specific, concise, and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here. If a specific interaction or moderation is important to your research, you can list that as a separate hypothesis.
   4.2. **Example**: If taste affects preference, then mean preference indices will be higher with higher concentrations of sugar.



https://osf.io/zab38/wiki/home/

Discussion: Pre-registration for fMRI

Problems:
- Too many decisions to make
- Analysis pipeline changes over time
- Power analysis

Some solutions and things to keep in mind:
- Template with overview of the many decisions to make: https://osf.io/dvb2e/
- Analysis plans can change: make amendments to pre-registration or describe differences in paper
- Power analysis for fMRI: http://neuropowertools.org; http://fmripower.org/

Resources: https://docs.google.com/document/d/1YrBc_bFlnWJVSjLjqQ_rRKtRMh9TTLLUcYMCsORg7Y0/edit

Data Management – A standard for data organization

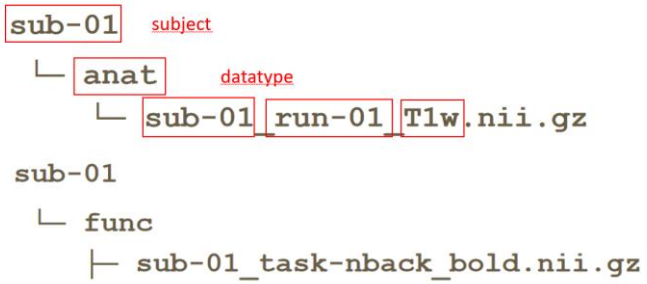
https://en.wikibooks.org/wiki/SPM/BIDS

Why is it important?

- Makes your own life easier (e.g. reproducing results, adapting scripts)
- Enhances understandibility of data structure and reproducility of results within own lab
- Facilitates sharing data and using publicly shared data

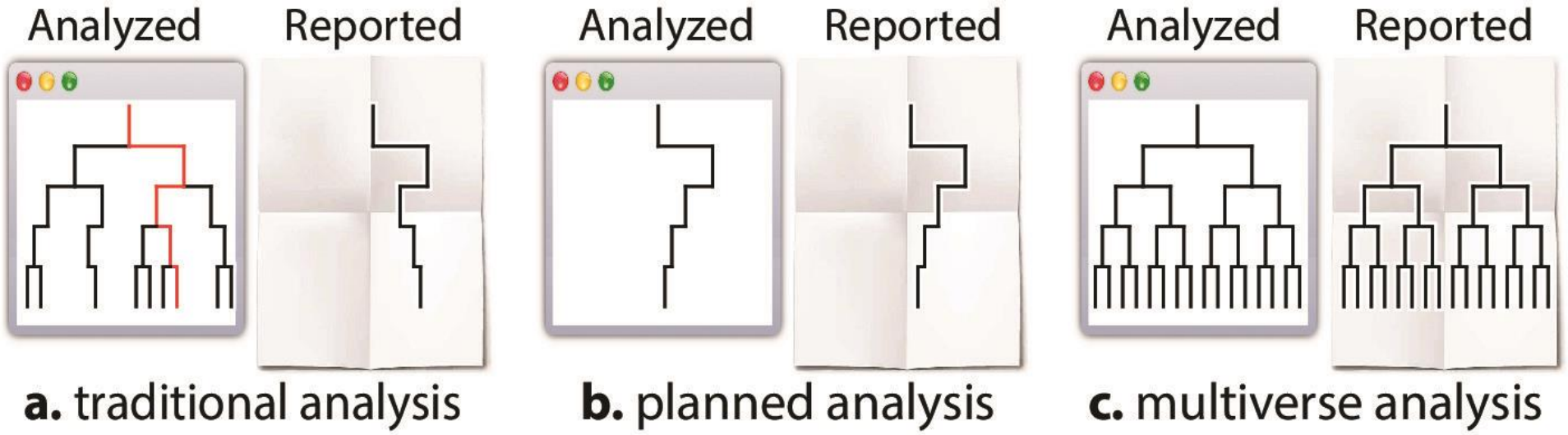
BIDS (as an example)

1) Json files: metadata and side-car file
2) TSV files: participant informations, events onsets etc
3) Nii-gz files. Zipped nifti files MRI data




```
dset/
├── dataset_description.json
├── participants.tsv
└── sub-01
    ├── anat
    │   ├── sub-01_T1w.json
    │   └── sub-01_T1w.nii.gz
    ├── dwi
    │   ├── sub-01_dwi.bval
    │   ├── sub-01_dwi.bvec
    │   ├── sub-01_dwi.json
    │   └── sub-01_dwi.nii.gz
    ├── fmap
    │   ├── sub-01_acq-dwi_dir-AP_epi.json
    │   ├── sub-01_acq-dwi_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-dwi_dir-PA_epi.json
    │   ├── sub-01_acq-dwi_dir-PA_epi.nii.gz
    │   ├── sub-01_acq-func_dir-AP_epi.json
    │   ├── sub-01_acq-func_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-func_dir-PA_epi.json
    │   └── sub-01_acq-func_dir-PA_epi.nii.gz
    └── func
        ├── sub-01_task-nback_run-01_bold.json
        ├── sub-01_task-nback_run-01_bold.nii.gz
        ├── sub-01_task-nback_run-01_events.tsv
        ├── sub-01_task-nback_run-01_sbref.json
        └── sub-01_task-nback_run-01_sbref.nii.gz
```

Data analysis: Multi verse



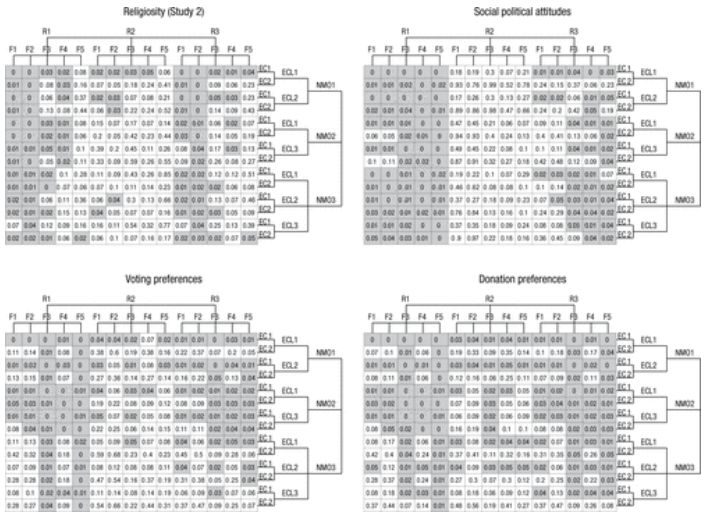
Increasing Transparency Through a Multiverse Analysis

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, more...
Show all authors

First Published September 29, 2016 | Research Article | Find in PubMed | Check for updates
https://doi.org/10.1177/1745691616658637
Article information
Altmetric 109

Abstract

Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could



Reproducible data analysis and results preparation

Why is it important?

- Enhances trust in our results
- Enhances efficiency (adjustments in the process, adaptation for new projects)
- Enables collaboration

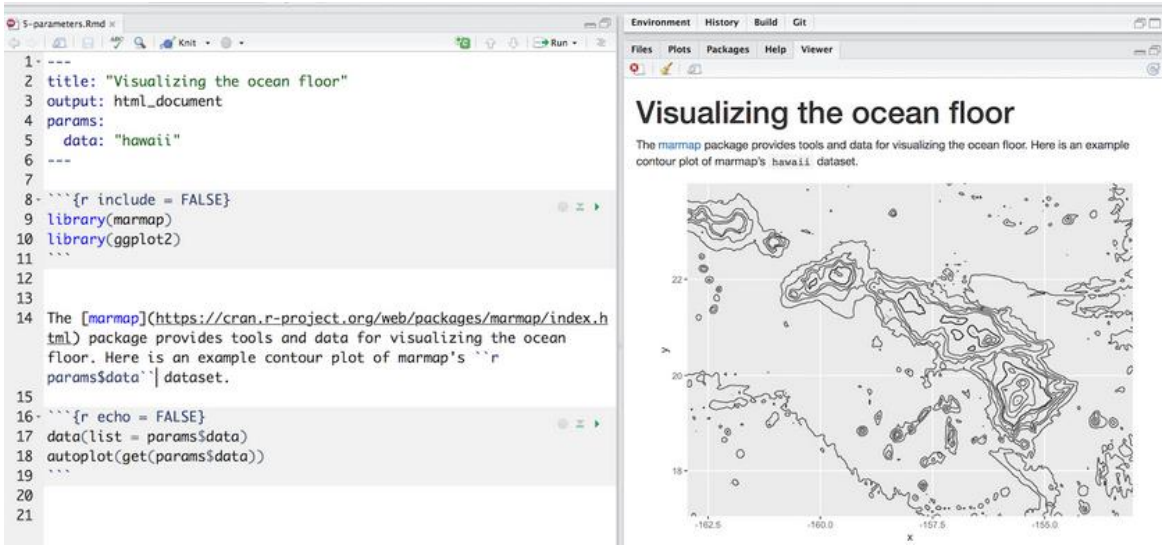
Aspiration: run single command to regenerate all results and figures from raw data

Recommendations for reproducible workflow:

- Consistent file structure and nomenclature
- Machine-readable file types
- Recording and controlling tools, libraries, versions etc.
- Automate as much as possible
- Version control

Reproducible data analysis: Tools

Data organization: custom sccript or converters



Reproducible manuscript writing

Why?

- Better control over results; reproducibility ensured
- Changes in analyses and results are implemented automatically
- Facilitates collaborative working on manuscript (analyses, figures, writing)

Papaja package using R Markdown:
https://github.com/crsh/papaja
Workshop: March, 29 and 30

```
# Material and methods

## Participants

```{r echo=FALSE, message=F, warning=F, results=F}

### Sample characteristics of T0:
# Compute n
n_T0 <- length(dataCovAll$id)

# Table gender, select females only
genderTable_T0 <- table(dataCovAll$gender)
females_T0 <- genderTable_T0[grep("female", names(genderTable_T0))]

# Age: Mean and SD, keep trailing zeros
age_mean_T0 <- format(round(mean(dataCovAll$age_T0, na.rm = T), 2), nsmall = 2)
age_sd_T0 <- format(round(sd(dataCovAll$age_T0, na.rm = T), 2), nsmall = 2)
age_range_T0 <- paste(range(dataCovAll$age_T0), collapse = " - ")

### Sample characteristics of T1:
# Compute n
n_T1 <- length(dataCovAll$id[which(dataCovAll$dropout == 0)])

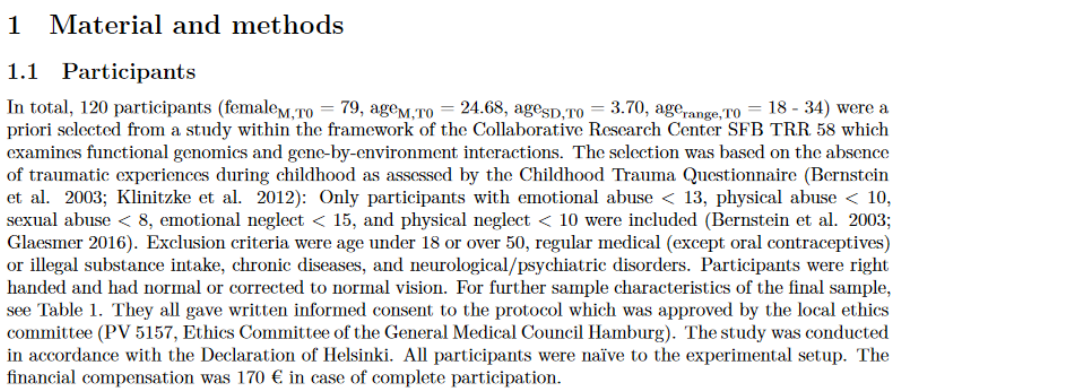
# Table gender, select females only
genderTable_T1 <- table(dataCovAll$gender[which(dataCovAll$dropout == 0)])
females_T1 <- genderTable_T1[grep("female", names(genderTable_T1))]

# Age: Mean and SD
age_mean_T1 <- format(round(mean(dataCovAll$age_T1[which(dataCovAll$dropout == 0)],
                                 na.rm = T), 2), nsmall = 2)
age_sd_T1 <- format(round(sd(dataCovAll$age_T1[which(dataCovAll$dropout == 0)],
                             na.rm = T) ,2), nsmall = 2)
age_range_T1 <- paste(range(dataCovAll$age_T1[which(dataCovAll$dropout == 0)],
                            na.rm = T), collapse = " - ")

### Compute n of complete cases
n_complete <- length(dataCovComplete$id)

```

In total, 120 participants (female~M,T0~ = `r females_T0`, age~M,T0~ = `r age_mean_T0`, age~SD,T0~ = `r age_sd_T0`, age~range,T0~ = `r age_range_T0`) were a priori selected from a study within the framework of the Collaborative Research Center SFB TRR 58 which examines functional genomics and gene-by-environment interactions. The selection was based on the absence of traumatic experiences during childhood as assessed by the Childhood Trauma Questionnaire [@bernstein2003; @klinitzke2012]: Only participants with emotional abuse \< 13, physical abuse \< 10, sexual abuse \< 8, emotional neglect \< 15, and physical neglect \< 10 were included [@bernstein2003; @glaesmer2016]. Exclusion criteria were age under 18 or over 50, regular medical (except oral contraceptives) or illegal substance intake, chronic diseases, and neurological/psychiatric disorders. Participants were right handed and had normal or corrected to normal vision. For further sample characteristics of the final sample, see Table \@ref(tab:Descript-table). They all gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). The study was conducted in accordance with the Declaration of Helsinki. All participants were naïve to the experimental setup. The financial compensation was 170 € in case of complete participation.
```



1 Material and methods

1.1 Participants

In total, 120 participants ($female_{M,T0} = 79$, $age_{M,T0} = 24.68$, $age_{SD,T0} = 3.70$, $age_{range,T0} = 18 - 34$) were a priori selected from a study within the framework of the Collaborative Research Center SFB TRR 58 which examines functional genomics and gene-by-environment interactions. The selection was based on the absence of traumatic experiences during childhood as assessed by the Childhood Trauma Questionnaire (Bernstein et al. 2003; Klinitzke et al. 2012): Only participants with emotional abuse < 13, physical abuse < 10, sexual abuse < 8, emotional neglect < 15, and physical neglect < 10 were included (Bernstein et al. 2003; Glaesmer 2016). Exclusion criteria were age under 18 or over 50, regular medical (except oral contraceptives) or illegal substance intake, chronic diseases, and neurological/psychiatric disorders. Participants were right handed and had normal or corrected to normal vision. For further sample characteristics of the final sample, see Table 1. They all gave written informed consent to the protocol which was approved by the local ethics committee (PV 5157, Ethics Committee of the General Medical Council Hamburg). The study was conducted in accordance with the Declaration of Helsinki. All participants were naïve to the experimental setup. The financial compensation was 170 € in case of complete participation.

Preprints I

What?

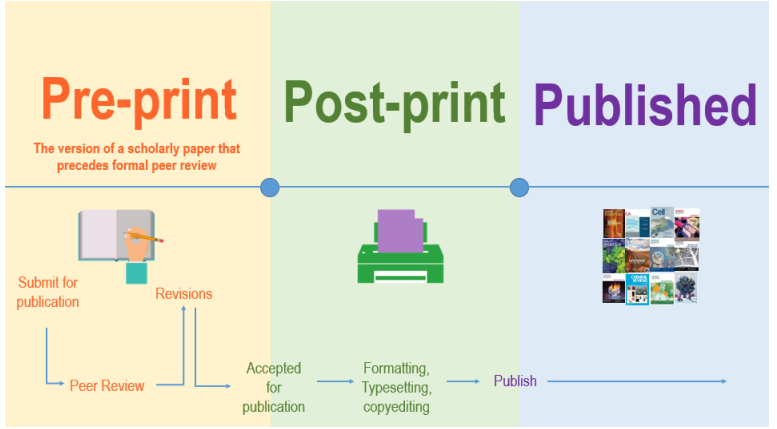
- Non-peer reviewed unpublished version of paper made available to the public (usually on preprint server)

Why?

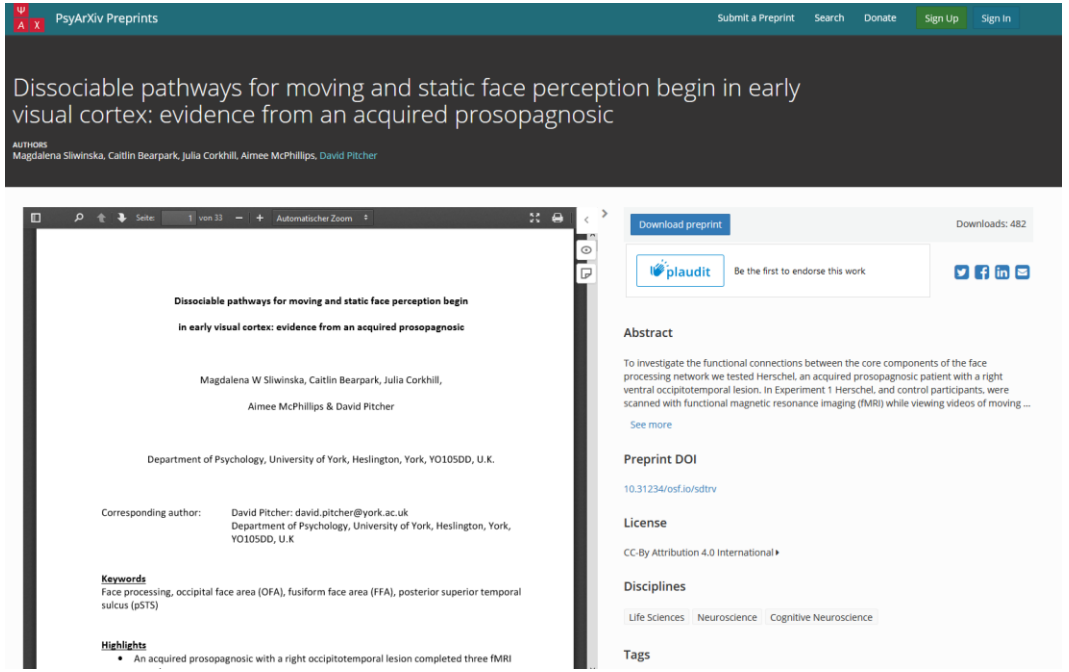
- Making new results available faster (avoid being scooped)
- Receive feedback outside of peer-review

How?

- Check guidelines of journal you are intending to submit to
- Choose preprint server and upload document
- Update postprint with link/reference to peer-reviewed version later

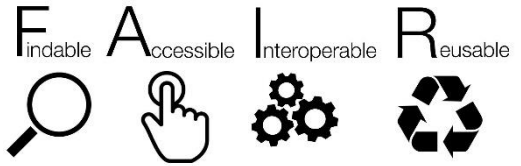


Preprints II



Questions and concerns about preprints?
Do you submit them?
Do you cite them?

FAIR principle



Findable:
- Use of unique and persistent identifed (doi)
- Machine-readable metadata
- Use of searchable repositiory

Accessible:
- Open and freely available data
- Can be retrieved by identifier

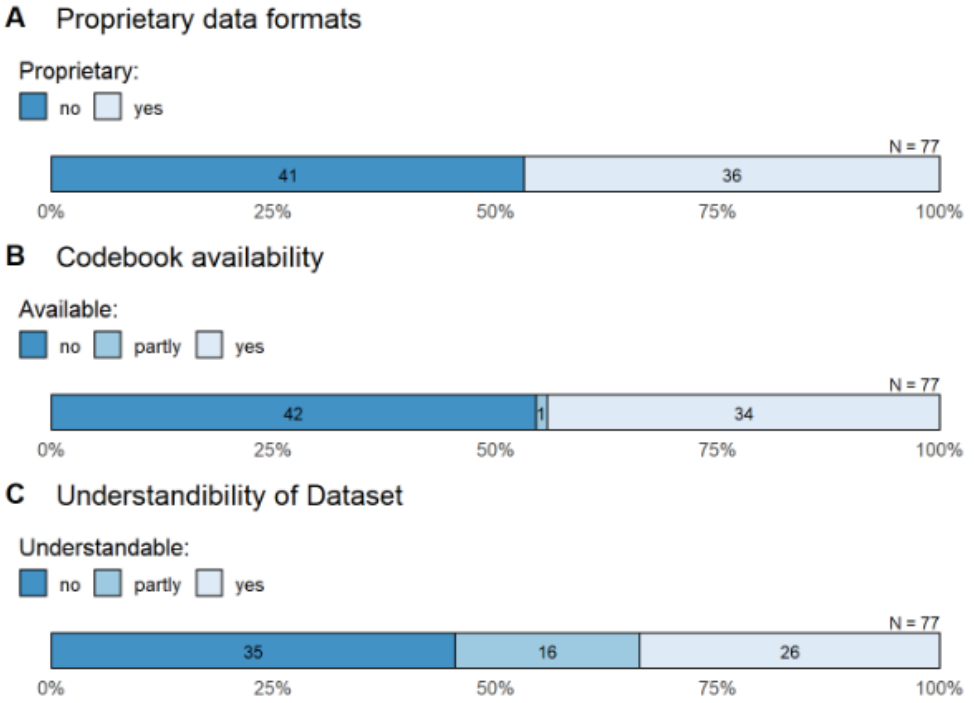
Interoperable
- Use of accessible and broadly applicable language for representation (i.e. avoid proprietary data formats)
- Use vocabulary following the FAIR principle

Reusable
- Data are described in detail
- Provide data usage license (e.g. CC-By Attribution 4.0 International)

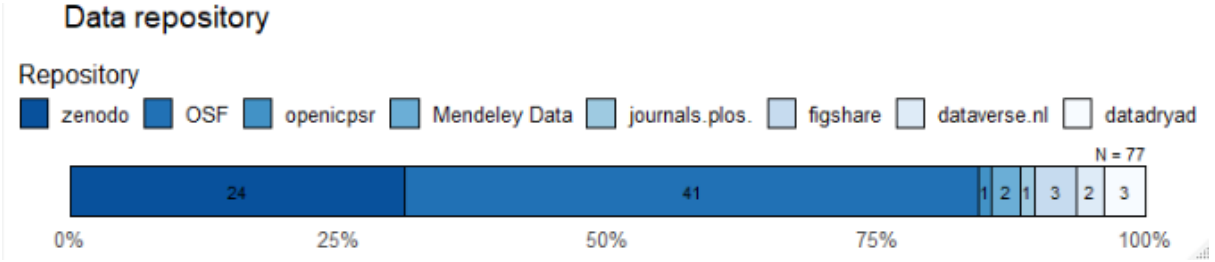
Wilkinson et al., 2018, *Scientific Data*

Data and Code Sharing: FAIR principle applied

- Create machine-readable metadata (xml, html, csv, json)
- Label your **variables** and provide a **codebook** (incl. description of variable names, data transformations, measurement units, measurement scales, questionnaire versions etc.)
- Save data files in non-proprietary formats (csv, tsv)
- Share primary data: first instance of digital data in standardized formats (e.g. EDF, NIFTI) (Recommendations by DGP, 2020)
- Scripts: comment a lot, make them easy to use for others
- Find the right **repository**
- Provide a **license**



Repositories



- Avoid personal or lab websites
- Avoid university of journal websites
- Chose repository that adheres to EU regulations, ensures persistent data acesss and link to identifier

Repository	**Country / Jurisdiction**	**Costs**	**Identifer (Y/N)**	**Machine Searchable (Y/N)**	**Usage statistics (Y/N)**
Zenodo	EU	free	Y	Y	y
OSF	USA	free	Y	N	Y
Mendeley	Netherlands	free	Y	N	Y
Figshare	UK	Free up to 100GB	Y	N	Y
Dryad	USA	$120/package	Y	N	Y
Dataverse	USA	free	Y	N	Y
ICPSR	USA	Free up to 2GB	Y	N	Y
GESIS	Germany	free	Y	N	Unclear

See also Klein et al., 2018, *Collabra: Psychology*

Licensing

Important to know:

- Important: primary data are not generally protected by copyright!
- Primary data is factual information and not a personal intellectual creation or expression of yourself
- But data that you collated, manipulated , transformed, etc. falls under the laws of copyright

Why?

- Informs people what they are allowed to do with your data.

Which license to choose?

- Easiest to use standard license
- CC (creative commons) license: places data in the public domain
- Most commonly used: CC by 4.0 International

This license lets others distribute, remix, adapt, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

Beyond FAIR: Ethical considerations

- General Data Protection Regulation (GDPR) passed in 2016
- Active consent for data usage
- Separate consent for usage of data within project and data for unknown purpose (if shared with others)
- Explicit consent for data sharing in and beyond repositories that adhere to GDPR standards (i.e. US based servers)
- (Pseudo)anonymization: identification of data subjects no longer possible (i.e. anonymization key should be destroyed when project is completed)
- Defacing of MRI data


What is important for
Data Protection
in science in the future?

General and specific changes in data protection for scientific use resulting from the EU General Data Protection Regulation

Katrin Schaar

July 2016

SPONSORED BY THE
Federal Ministry of Education and Research

For practical tips: Meyer et al., 2017
https://doi.org/10.1177/2515245917747656

https://www.konsortswd.de/wp-content/uploads/RatSWD_WP_258.pdf