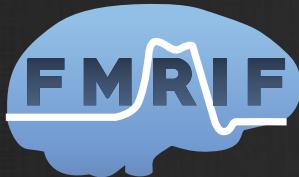


# Data Sharing and Open Science in Neuroimaging

Adam Thomas

Data Science and Sharing Team, FMRIF, NIMH



# Credits

Material borrowed, adapted, and/or stolen from:

- Russ Poldrack



- Chris Gorgolewski



- Brian Nosek



- Tal Yorkoni



- Niko Kriegeskorte



- Tom Nichols



- Phil Bourne



# Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

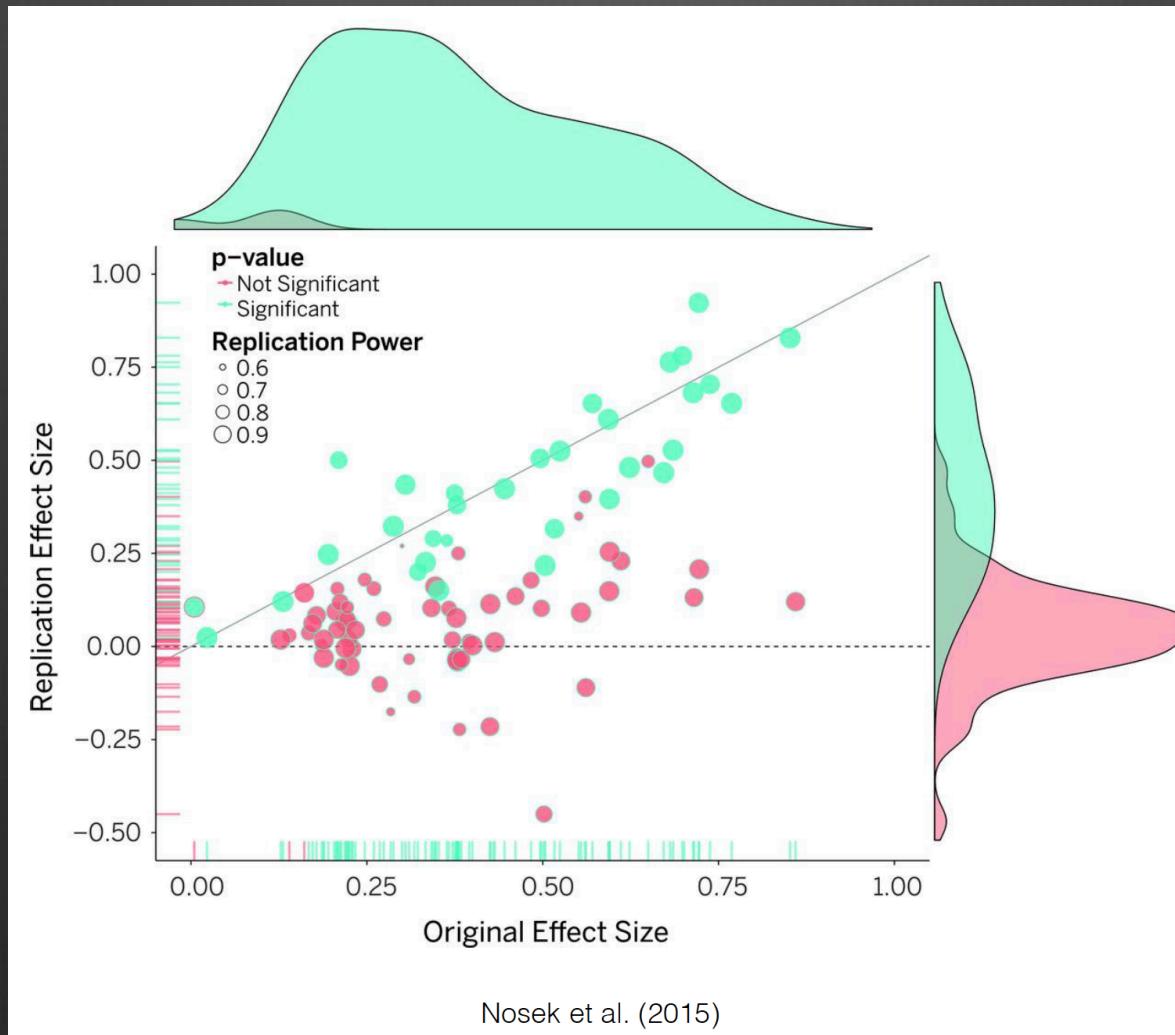
# Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

# Outline

- The Problem
- What is Open Science?
- How do I do Open Science?

# Problem: Reproducibility



# The Problem: Reproducibility



PLoS Medicine | www.plosmedicine.org

August 2005 | Volume 2 | Issue 8 | e124

## 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 of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

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 confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

**It can be proven that most claimed research findings are false.**

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 is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is  $R/(R + 1)$ . The probability of a study finding a true relationship reflects the power  $1 - \beta$  (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate,  $\alpha$ . Assuming that  $c$  relationships are being probed in the field, the expected values of the  $2 \times 2$  table are given in Table 1. After a research finding has been claimed based on

# Problem: Wasted time & resources



“How much time do you spend handling, reorganizing, and managing your data as opposed to actually *doing* science?”

- Median answer is 80%

# Problem: Wasted Time & resources

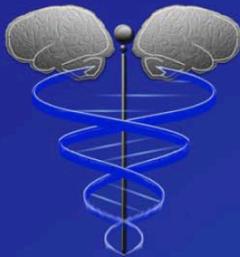
## Unpublished Data

- File drawer problem
- Lost staff & lost metadata
- Underutilized data



# The Problem

Lack of transparency and reproducibility hinders integration



# NIMH

National Institute of Mental Health  
Intramural Research Program (IRP)

*Blue Ribbon Panel*

June 6, 2008

*Final Report*

“The Blue Ribbon Panel proposes that basic and clinical groups in NIMH IRP be linked more closely than is generally the case in universities. Linking basic and clinical teams of investigators may facilitate the translational goals of understanding disease mechanisms and developing novel therapies.“

# The Problem... is not new

## Research in the Service of Mental Health

Summary Report of the Research Task Force of the National Institute of Mental Health

A comprehensive and detailed report of the NIMH Research Task Force, totaling over 400 pages, is for sale by the Superintendent of Documents, Government Printing Office, Washington, D.C. 20402. Order DHEW Publication No. (ADM) 75-236 Printed 1975

### 3. *The Need to Broaden the Use of Research Findings*

The greatest single need in this area is an explicit policy on which to base an Institute-wide effort to disseminate research findings, and, whenever appropriate, to foster their use.

### 4. *The Need for Synthesis and Integration*

There has been a natural tendency to use research funds mainly for the development of new knowledge. Relatively neglected has been the need to bring together and evaluate findings in a given area, consider them in relation to findings from other mental health research areas, and determine the implications for further research and for application. NIMH should recognize that the synthesis and integration of research results may often be as important as the research itself.

“Relatively neglected has been the need to bring together and evaluate findings in a given area and consider them in relation to findings from other mental health research areas [...] NIMH should recognize that the synthesis and integration of research results may often be as important as the research itself”

- Research Task Force of the NIMH, 1975

# The Problem... is not new

<https://archive.org/details/nimh-nihlibrary>

# National Institute of Mental Health (NIMH) Publications

Annual reports and publications created by the National Institute of Mental Health (NIMH) have been digitized and made freely available on the Web as a service of the NIH Library.

About

Collection

SORT BY RELEVANCE · VIEWS · TITLE · DATE PUBLISHED · CREATOR

Contagious and infectious diseases among the Indians

1913

194 | 0 | 0 | 0

Vol 1967-68 v.1-2: Report of program activities :

1954

903 | 0 | 0 | 0

Vol 1972 pt.1-2: Report of program activities :

1954

990 | 0 | 0 | 0

Vol 1955: Report of program activities :

1954

1,094 | 0 | 0 | 0

Vol 1956: Report of program activities :

1954

899 | 0 | 0 | 0

Vol 1960: Report of program activities :

1954

646 | 0 | 0 | 0

Vol 1971 pt.1-2: Report of program activities :

1954

1,122 | 0 | 0 | 0

Vol 1954: Report of program activities :

1954

643 | 0 | 0 | 0

Vol 1957: Report of program activities :

1954

1,139 | 0 | 0 | 0

Vol 1959 v.1: Report of program activities :

1954

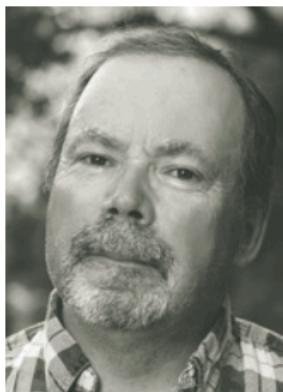
846 | 0 | 0 | 0

# Problems: The big-data revolution

## PERSPECTIVE

### Sustaining the big-data ecosystem

*Organizing and accessing biomedical big data will require quite different business models, say Philip E. Bourne, Jon R. Lorsch and Eric D. Green.*



**B**iomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-

recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

**FAIR AND EFFICIENT**

OPEN SCIENCE:

WHY



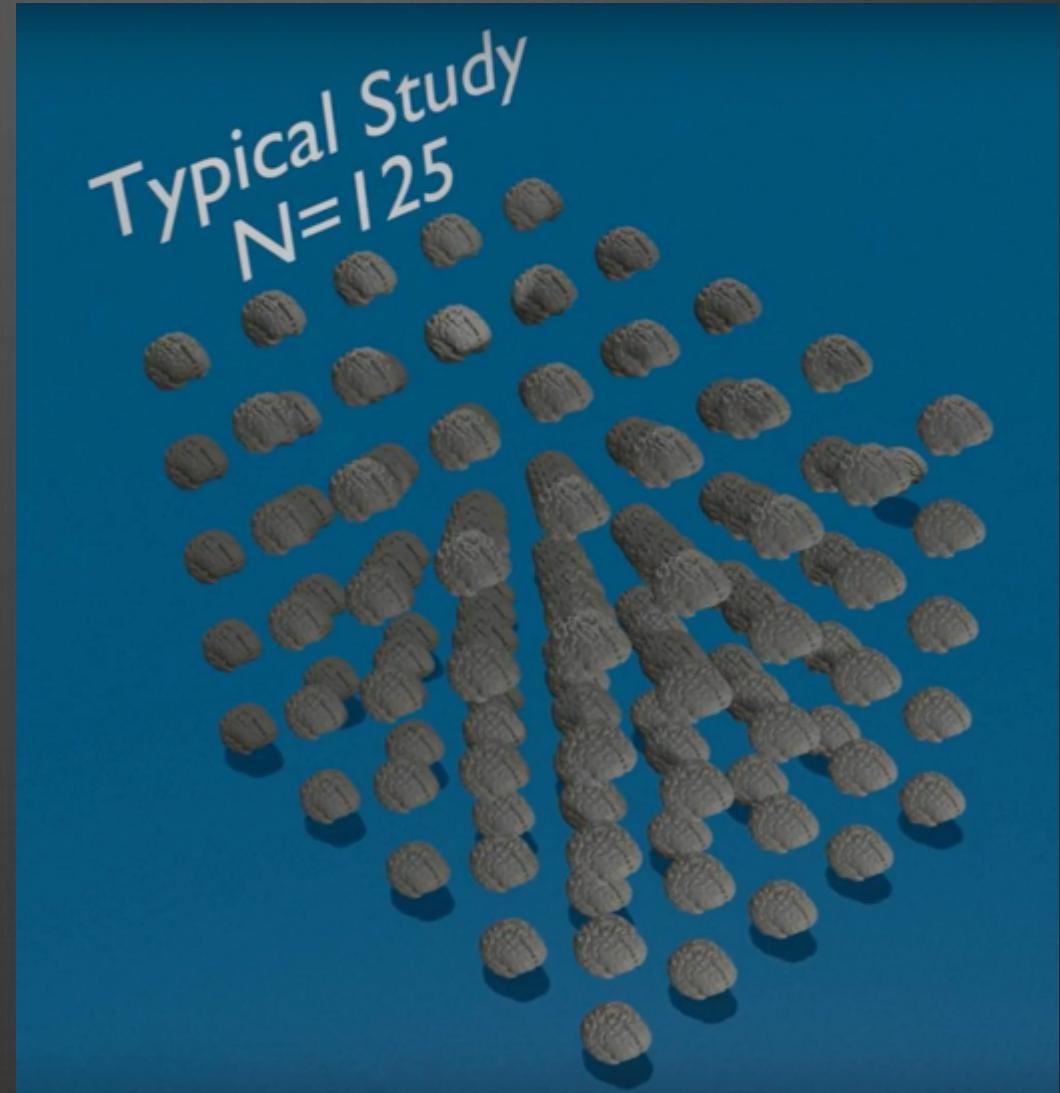
WHAT



HOW

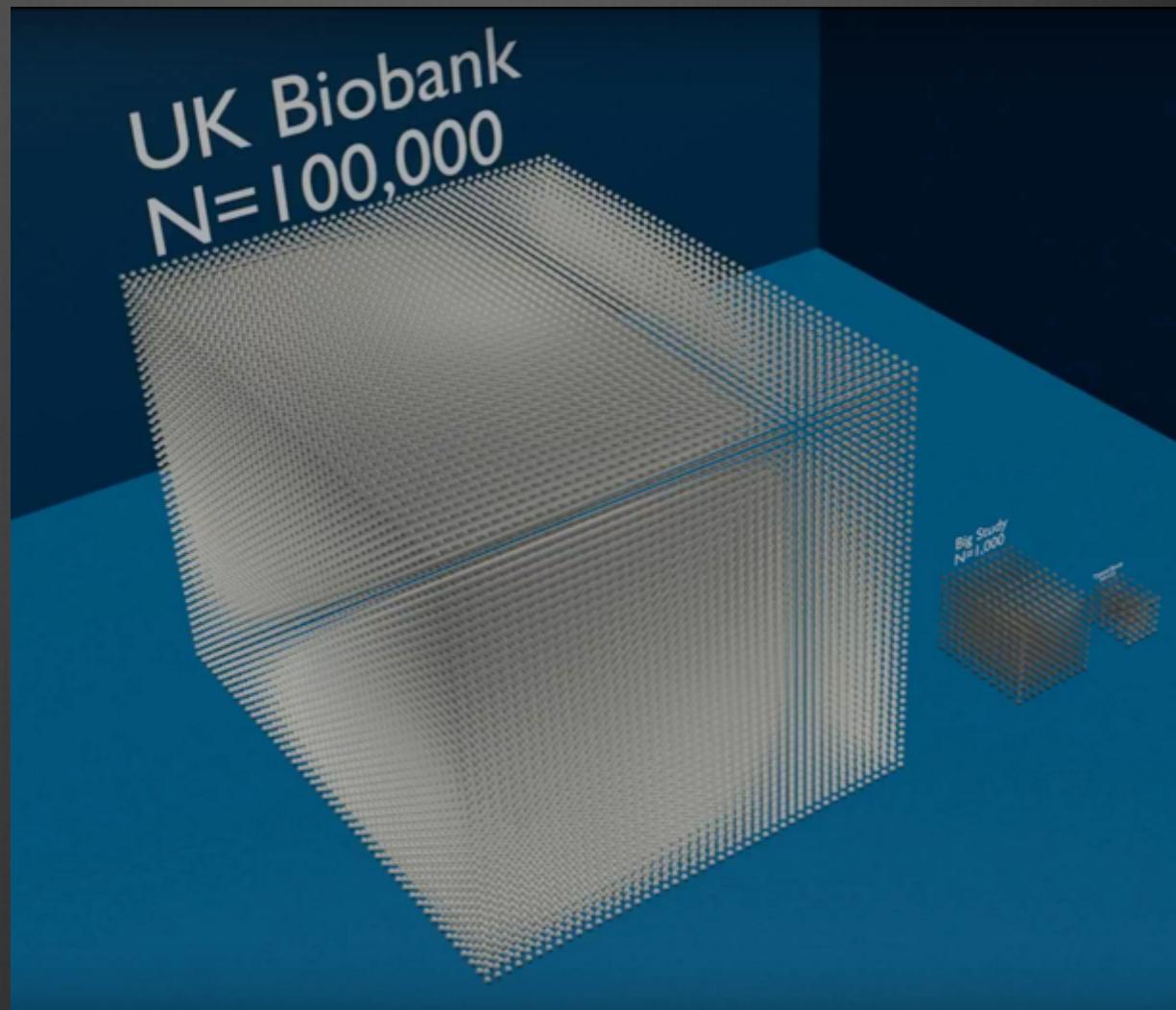
# Problems: The big-data revolution

UK  
Biobank  
Imaging  
Initiative



# UK Biobank Imaging Initiative

## Problems: The big-data revolution



# Problems: The big-data revolution



Obama's precision medicine initiative will aim to enroll a large number of people in a genetic database representing the U.S. population.

Amy West/Flickr (CC BY 2.0)

## President Obama's 1-million-person health study kicks off with five recruitment centers

By Jocelyn Kaiser | Jul. 7, 2016 , 5:00 PM

# Problems: The big-data revolution

FILM

vs.

DIGITAL



OPEN SCIENCE:

WHY



WHAT



HOW

# Outline

- The Problem
  - Reproducibility
  - Wasted resources
  - Lack of integration
  - Ill-prepared to work with big datasets
- What is Open Science?
- How do I do Open Science?

# Outline

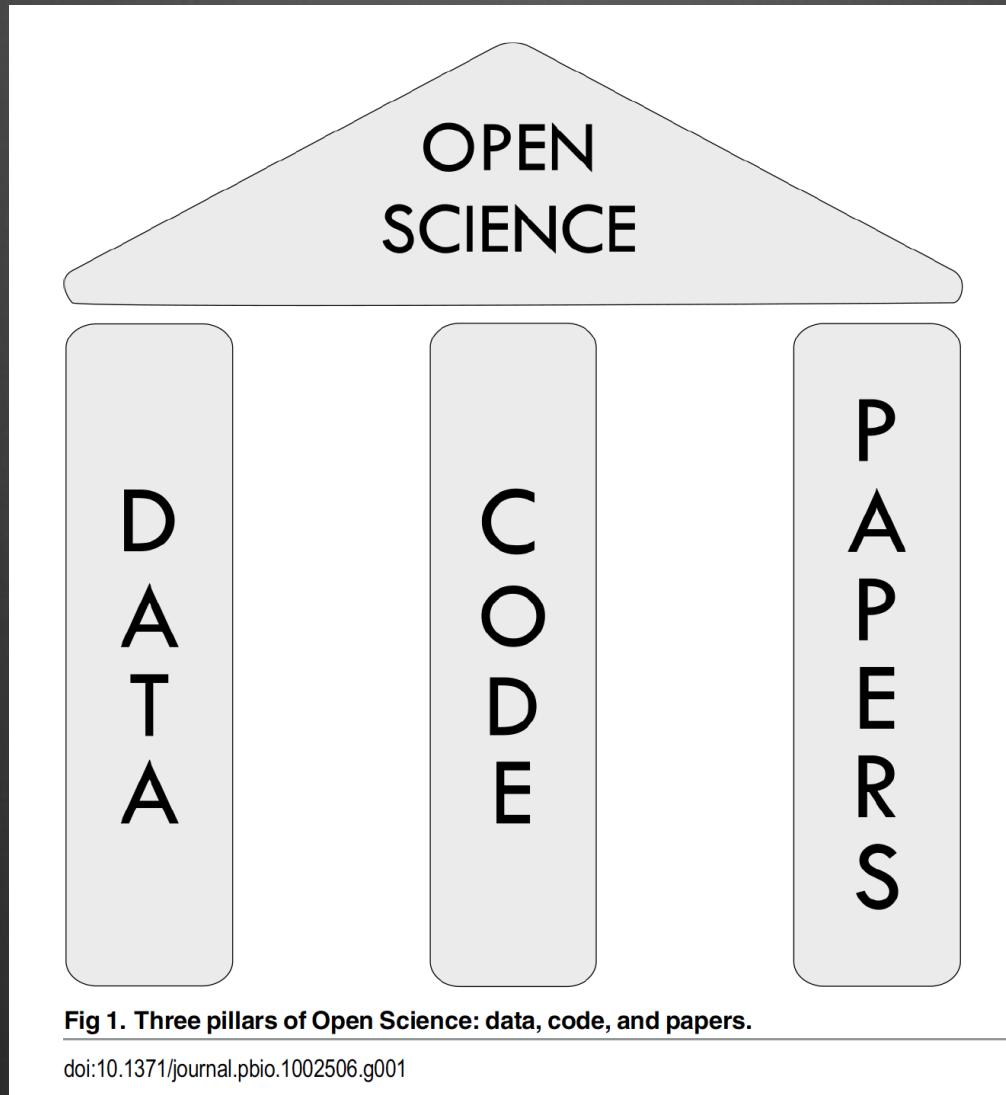
- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

# A manifesto for reproducible science

Marcus R. Munafò<sup>1,2\*</sup>, Brian A. Nosek<sup>3,4</sup>, Dorothy V. M. Bishop<sup>5</sup>, Katherine S. Button<sup>6</sup>, Christopher D. Chambers<sup>7</sup>, Nathalie Percie du Sert<sup>8</sup>, Uri Simonsohn<sup>9</sup>, Eric-Jan Wagenmakers<sup>10</sup>, Jennifer J. Ware<sup>11</sup> and John P. A. Ioannidis<sup>12,13,14</sup>

- The Problem
- Methods
  - Cognitive Bias
  - Methodological Training
  - Independent Method support (and oversight)
  - Encouraging Team Science
- Reporting and dissemination
  - Pre-registration
  - Quality of Reporting (checklist & guidelines)
- Reproducibility
  - Transparency
  - Data Sharing
- Evaluation
  - Peer Review
- Incentives
  - Changing cultural norms
  - Badges

# What is Open Science?



OPEN SCIENCE:

WHY

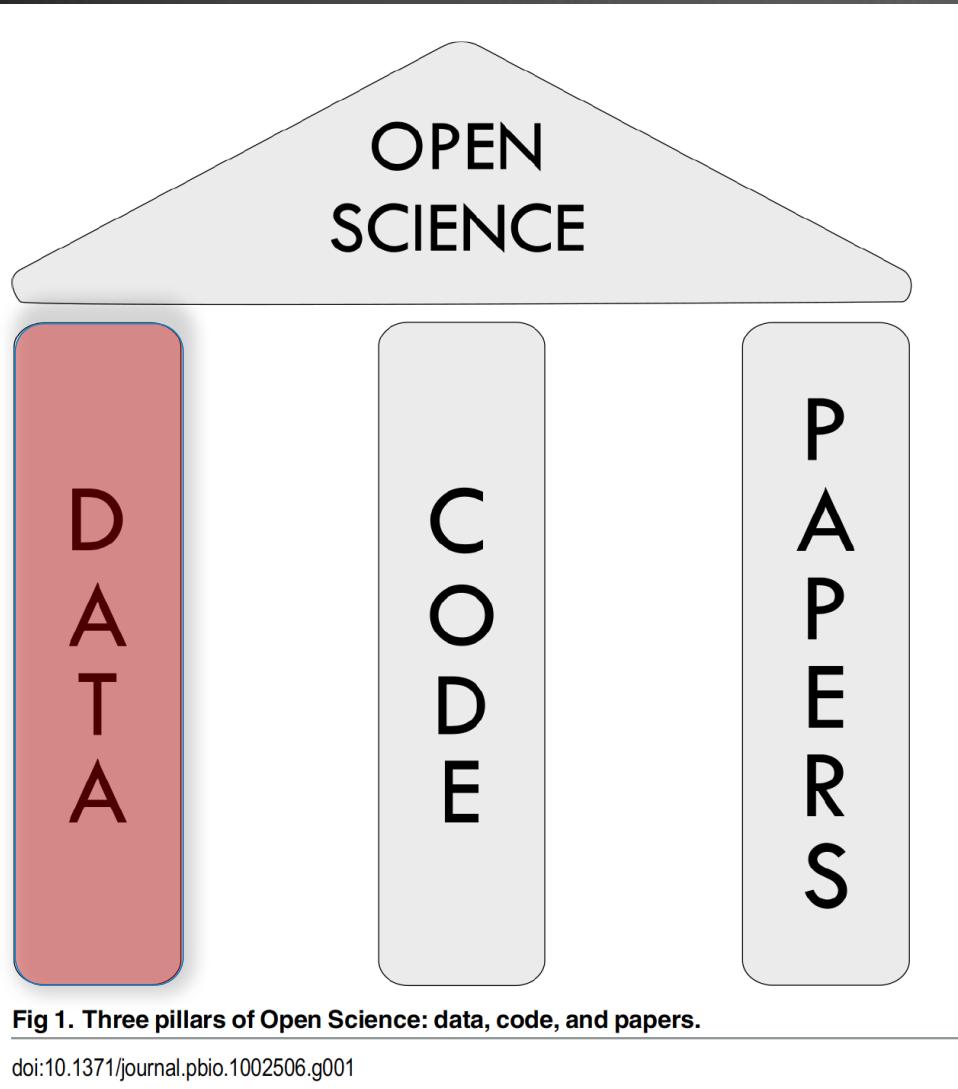


WHAT



HOW

# What is Open Data?



Data deposited in a public, community-recognized repository with a stable DOI

Follows FAIR Principle

- Findable
- Accessible
- Intra-operable
- Reusable

Should be deposited *before* publication

# Open Data: Community recognized Repositories

## MRI Specific Repos

- OpenfMRI
- COINS
- FCP/INDI
- LONI
- LORIS
- NITRC
- XNAT Central
- ANIMA\*
- BALSA\*
- Neuovault\*

\* Statistical & derived data only

## Data Agnostic Repos

- FigShare
- Dryad
- DataVerse
- Open Science Framework
- NIMH Data Archive

Coming soon: A dedicated MRI image repository for MRI studies conducted at intramural NIMH

# What is Open Code?

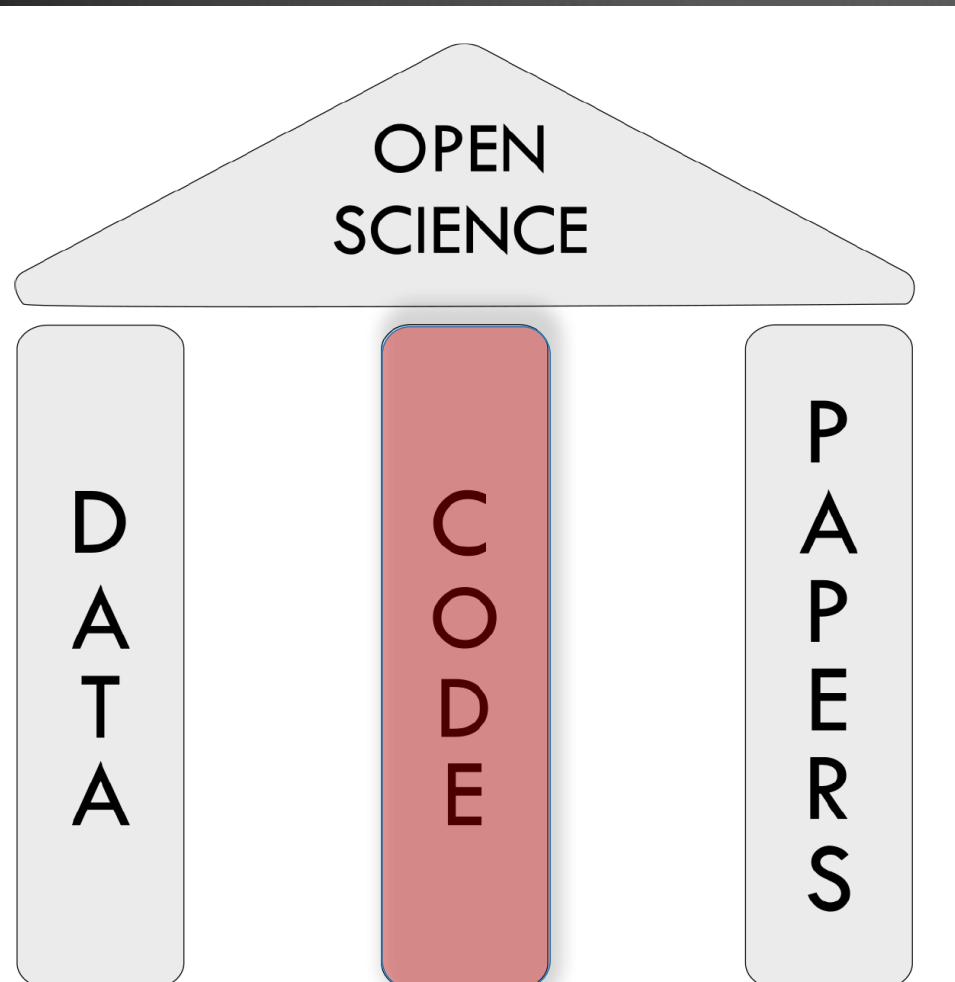
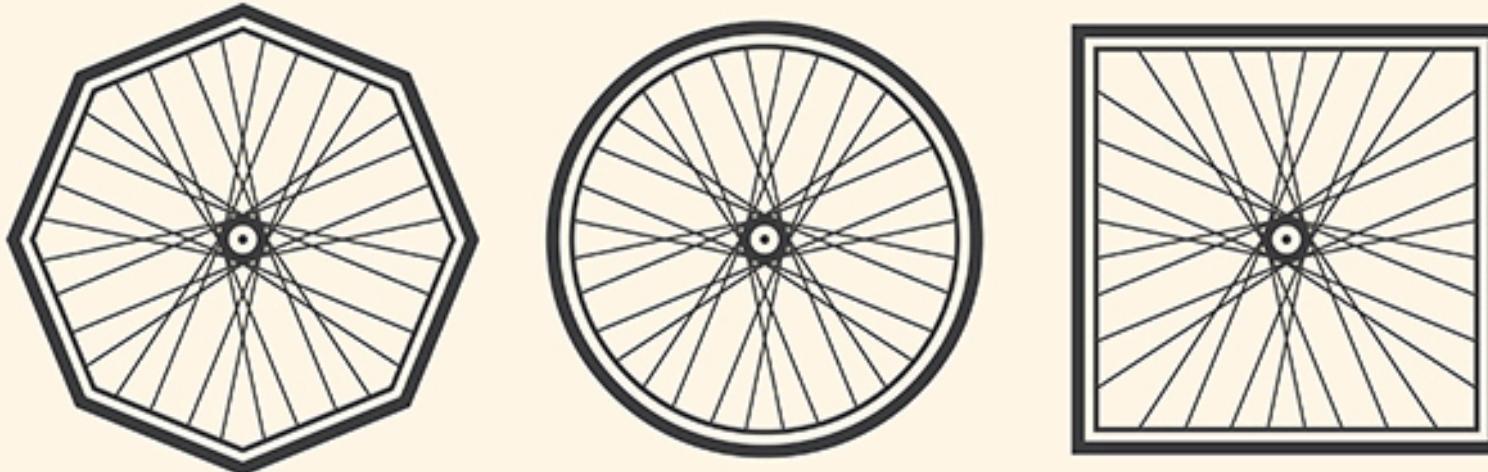


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

Open code enables greater reproducibility (includes non-code methods)

# Open Code – Don't Reinvent



Reuse and improve



**OPEN SCIENCE:**

**WHY**



**WHAT**



**HOW**

# Open Code - Version Control

Version control systems allows you to:

- Store all of your analysis in a central repository
- Keep a history of “snapshots” of your evolving analysis
- Quickly switch between different versions of your analysis
- Adopt and modify code from other scientists
- Collaborate



# What are Open Papers?

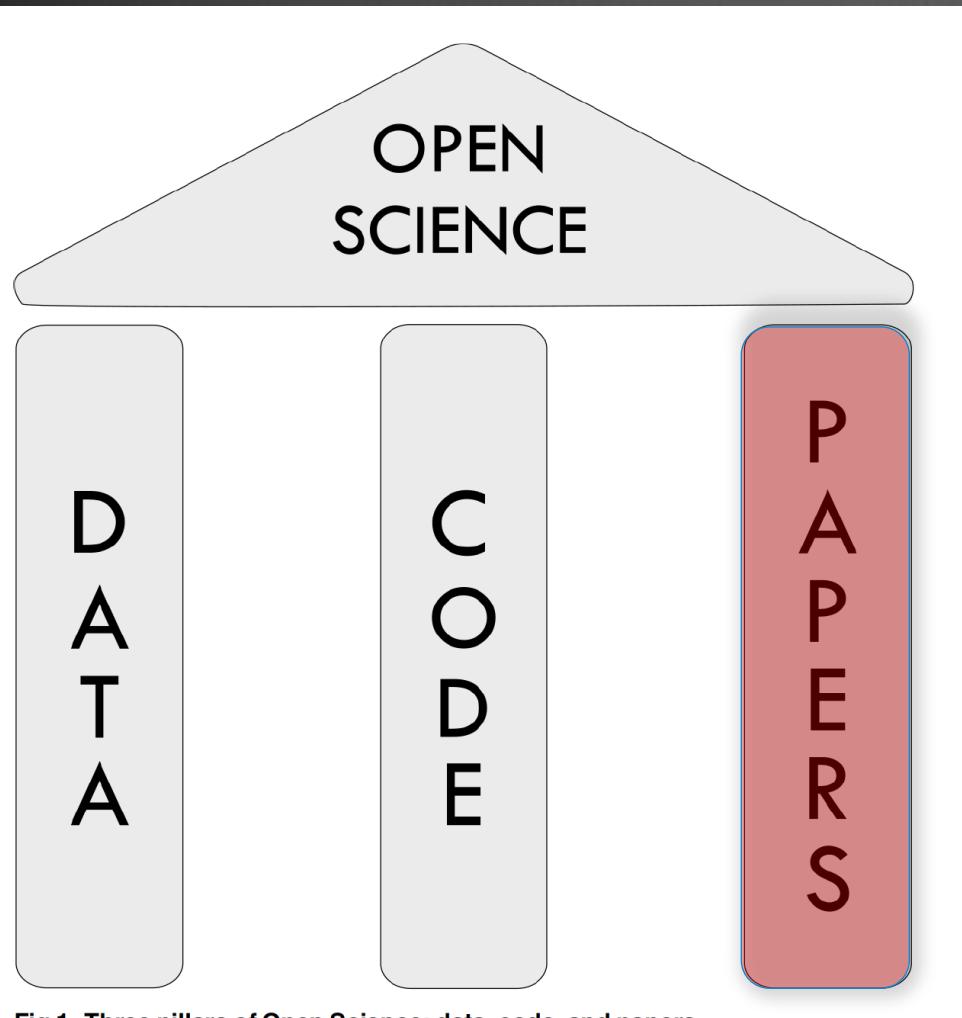


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

- Preprint posting
- Open access
- Open review

OPEN SCIENCE:

WHY



WHAT



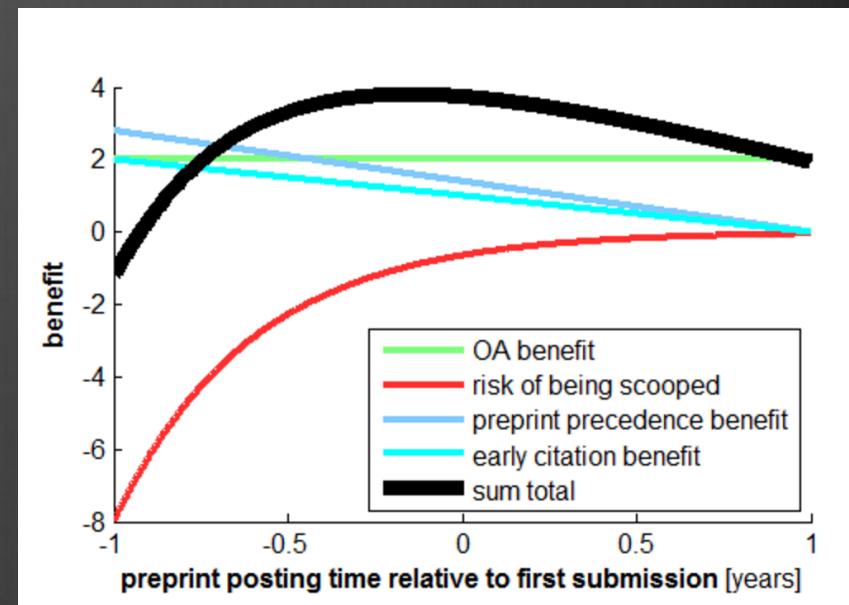
HOW

# Open Papers: Preprint posting

arXiv.org

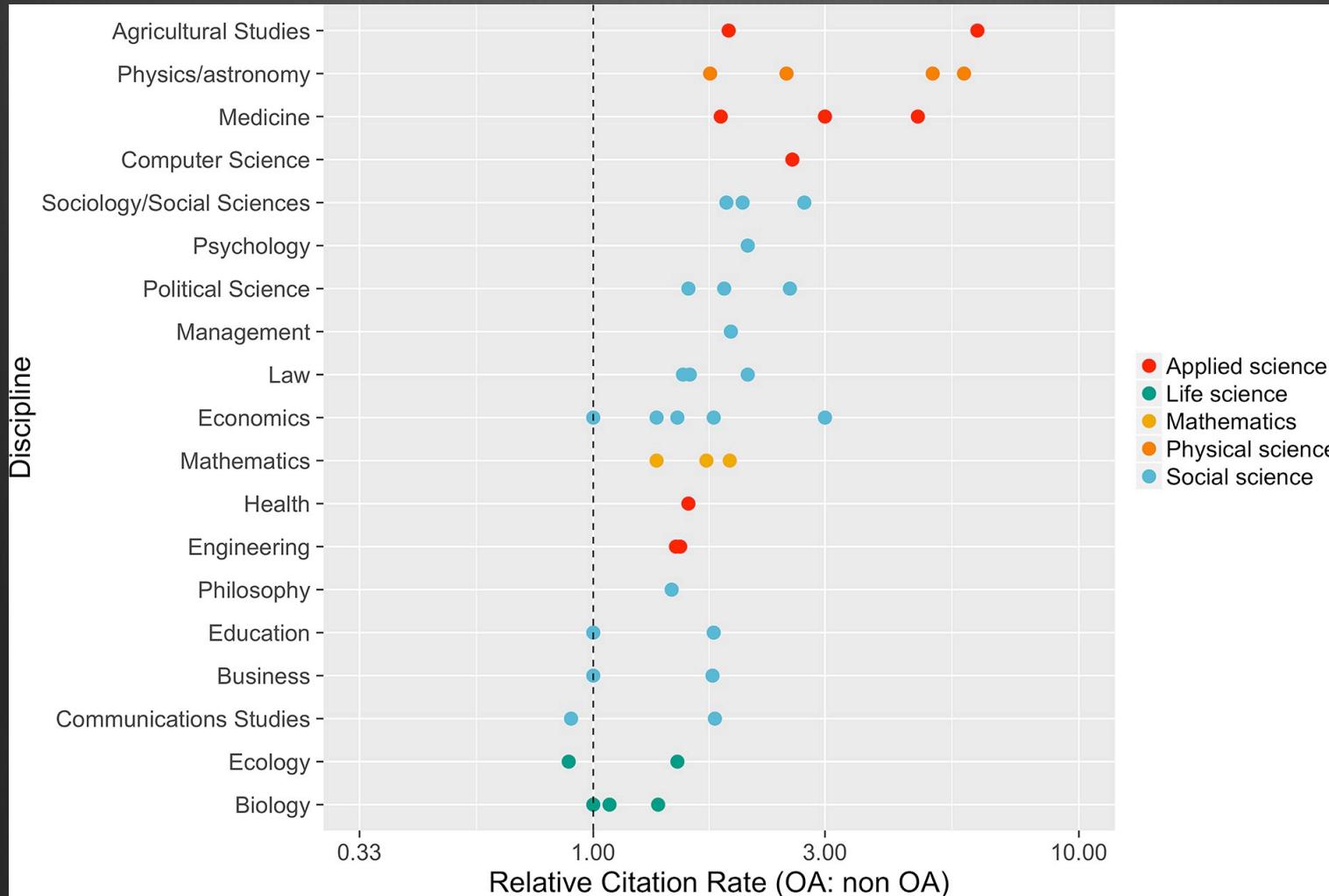
bioRxiv  
beta  
THE PREPRINT SERVER FOR BIOLOGY

- Benefits:
  - Open access
  - Catch errors
  - Earlier citation
  - Earlier precedence,  
prevent scooping
  - Speed and improve final submission



# Open Access

## Open access publication are cited more



<https://elifesciences.org/content/5/e16800%20>

mean citation rate of OA articles divided by mean citation rate of non-OA articles

# Open Review

PubPeer

The online journal club

 Search by DOI, PMID, arXiv ID, keyword, author, etc.

The PubPeer database contains all articles. Search results return articles with comments.  
To leave a new comment on a specific article, paste a unique identifier such as a DOI, PubMed ID, or arXiv ID into the search bar.

Search Publications

*the*  
**WINNOWER**

*The Winnower is founded on the principle that all ideas should be openly discussed, debated, and archived.*

- Public discussion of pros and cons of submission
- Optional anonymity
- Prevent low-quality and or biased review

OPEN SCIENCE:

WHY

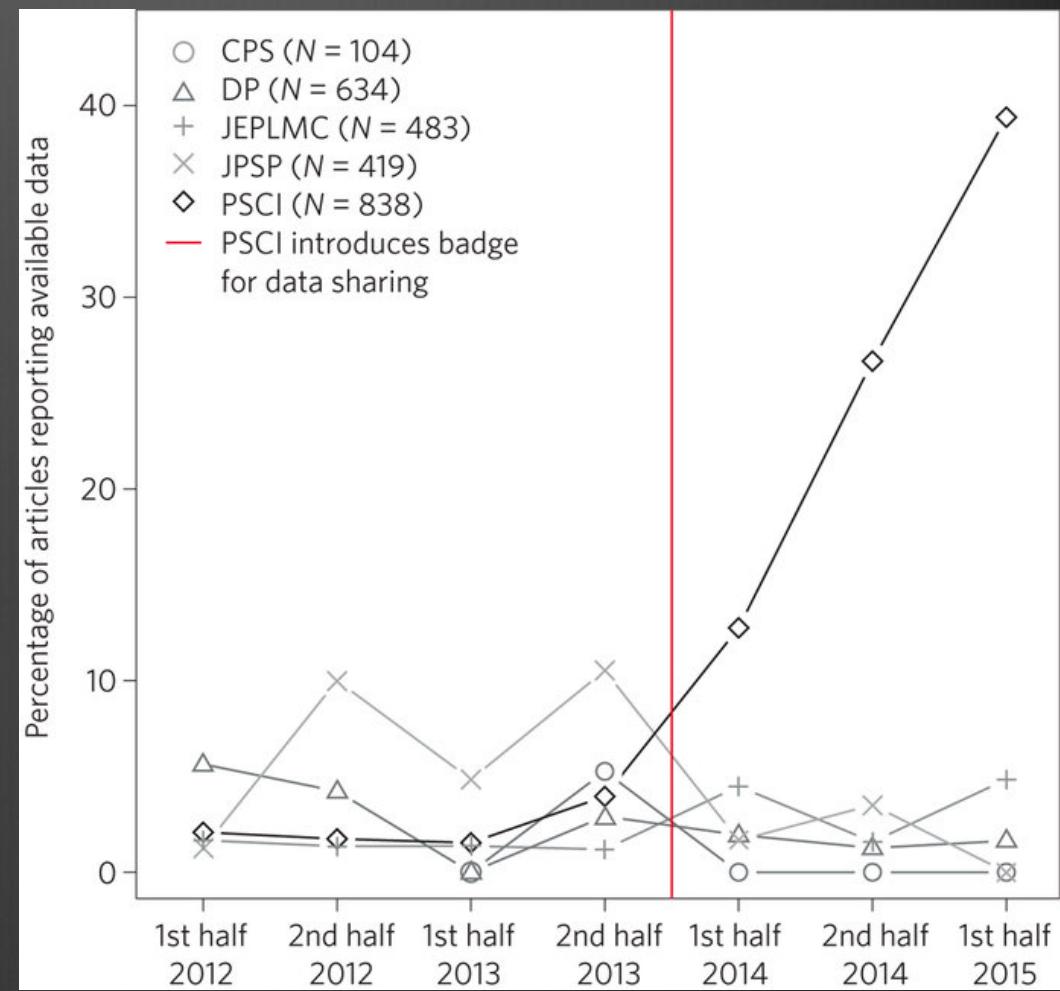


WHAT



HOW

# Incentives: Badges



OPEN SCIENCE:

WHY →

WHAT →

HOW

# Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

# How – Plan Ahead

- Get data sharing in your protocol:
  - NIMH Data Sharing Committee
  - <https://open-brain-consent.readthedocs.io>
- When designing, collecting, and analyzing consult with standards documents:
  - Enhancing Quality and Transparency of Health Research (EQUATOR) <http://www.equator-network.org>
  - Best Practices in Data Analysis and Sharing in Neuroimaging using MRI (COBIDAS) <http://dx.doi.org/10.1101/054262>

Open Brain Consent

latest

# Standards – EQUATOR & COBIDAS

- EQUATOR: Different standards for different designs
  - RCT, crossover, observational, etc.
- COBIDAS Sections
  1. Experimental Design
  2. Image Acquisition
  3. Preprocessing
  4. Statistical Modeling
  5. Results
  6. Data Sharing
  7. Reproducibility
- Both EQUATOR and COBIDAS focus on reporting,
- Reviewing them in advance will help you plan and design your study
- Also useful reference when reviewing papers

# Standards – EQUATOR & COBIDAS

## Checklists



### CONSORT 2010 checklist of information to include when reporting a randomised trial\*

Section/Topic	Item No	Checklist item	Reported on page No
<b>Title and abstract</b>			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
<b>Introduction</b> Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
<b>Methods</b> Trial design	3a	Description of trial design (such as parallel, factorial, etc.)	
	3b	Important changes to methods after trial commencement (for example,停止, addition of new interventions, or changes to study team)	
<b>Participants</b>	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
<b>Interventions</b>	5	The interventions for each group with their key features	
		Actual interventions administered if they differ from intended interventions	
<b>Outcomes</b>	6a	Completely defined pre-specified primary and any other key outcomes	
	6b	Any changes to trial outcomes after trial commencement	
<b>Sample size</b>	7a	How sample size was determined	
	7b	When applicable, explanation of any changes in study power and sample size	
<b>Randomisation:</b>			

Table D.1. Experimental Design Reporting

Aspect	Notes	Mandatory
<b>Number of subjects</b>	<i>Elaborate each by group if have more than one group.</i>	
Subjects approached		N
Subjects consented		N
Subjects refused to participate	Provide reasons.	N
Subjects excluded	Subjects excluded after consenting but before data acquisition; provide reasons.	N
Subjects participated and analyzed	Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.	Y
<b>Inclusion criteria and descriptive statistics</b>	<i>Elaborate each by group if have more than one group.</i>	
Age	Mean, standard deviation and range.	Y
Sex	Absolute counts or relative frequencies.	Y
Race & ethnicity	Per guidelines of NIH or other relevant agency.	N

# COBIDAS – Highlights

- Report scan parameters by exporting exam cards
- Preprocessing include *all* steps applied to the data before and must be reported
- For maximal transparency, report all regions of interest (ROIs) and/or experimental conditions examined as part of the research, so that the reader can gauge the degree of any HARKing
  - Hypothesizing After The Results are Known
  - It's OK to explore your data, just be clear that that is what you're doing

# COBIDAS – 4. Statistical Modeling

- Different reporting standards for different statistical approaches (univariate, multivariate, connectivity).
- Appendix C: Translates software used to statistical language

## Appendix C. Short descriptions of fMRI models

While any analysis software consists of myriad modelling decisions, an author must be able to describe the key facets of an analysis in the methods section of their paper. To facilitate this, and to suggest a level of detail that is useful to readers unfamiliar with the software yet not distractingly long, we provide short descriptions for the most commonly used statistical models in widely used software packages.

### C1. Task fMRI

Summaries for AFNI<sup>44</sup>, Freesurfer<sup>45</sup>, FSL<sup>46</sup>, & SPM<sup>47</sup> are based on versions AFNI\_2011\_12\_21\_1014, FreeSurfer 5.3, FSL 5.0.8 and SPM 12 revision 6470, respectively.

**AFNI 1<sup>st</sup> level – 3dDeconvolve:** Linear regression at each voxel, using ordinary least squares, drift fit with polynomial.

**AFNI 1<sup>st</sup> level – 3dREMLfit:** Linear regression at each voxel, using generalised least squares with a voxel-wise ARMA(1,1) autocorrelation model, drift fit with polynomial.

**AFNI 2<sup>nd</sup> level – 3dTtest:** Linear regression at each voxel, using ordinary least squares.

**AFNI 2<sup>nd</sup> level – 3dMEMA:** Linear mixed effects regression at each voxel, using generalized least squares with a local estimate of random effects variance.

**AFNI 2<sup>nd</sup> level – 3dMVM:** Multivariate ANOVA or ANCOVA at each voxel.

**AFNI 2<sup>nd</sup> level – 3dLME:** General linear mixed-effects modeling at each voxel, with separate specification of fixed and random variables.

**Freesurfer 1<sup>st</sup> Level – selxavg3-sess:** Linear regression at each surface element, using generalized least squares with a element-wise AR(1) autocorrelation model, drift fit with polynomial.

**Freesurfer 2<sup>nd</sup> Level – mri\_glmfit:** Linear regression at each surface element, using ordinary least squares.

# COBIDAS – 5. Results

- Mass univariate analyses should report:
  - All Effects tested
  - Tables of brain coordinates,
  - Thresholded maps
  - Extracted data
- Functional Connectivity
- Multivariate and Predictive Analysis
  - Evaluation : Quality of predicted fit
  - Interpretation: What does it mean

# COBIDAS – 6. Data Sharing

- Planning: “Data sharing is most onerous when done as an afterthought”
- “a comprehensive data management plan—that involves all authors, collaborators, funding agencies, and publishing entities—is essential”

# COBIDAS – 7. Reproducibility

Archiving: Think long term

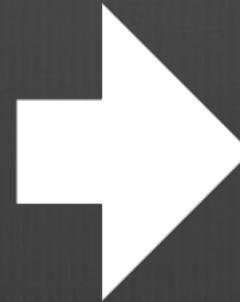
- Open-source software is more likely to be available long term
- URLs “decay” over time. Use Digital Object Identifiers (DOI) instead

# Organizing your data - BIDS

A simple and intuitive way to organize and describe your neuroimaging and behavioral data.

<http://bids.neuroimaging.io>

- 📁 dicomdir/
  - 📁 1208200617178\_22/
    - 📄 1208200617178\_22\_8973.dcm
    - 📄 1208200617178\_22\_8943.dcm
    - 📄 1208200617178\_22\_2973.dcm
    - 📄 1208200617178\_22\_8923.dcm
    - 📄 1208200617178\_22\_4473.dcm
    - 📄 1208200617178\_22\_8783.dcm
    - 📄 1208200617178\_22\_7328.dcm
    - 📄 1208200617178\_22\_9264.dcm
    - 📄 1208200617178\_22\_9967.dcm
    - 📄 1208200617178\_22\_3894.dcm
    - 📄 1208200617178\_22\_3899.dcm
  - 📁 1208200617178\_23/
  - 📁 1208200617178\_24/
  - 📁 1208200617178\_25/



- 📁 my\_dataset/
  - 📄 participants.tsv
  - 📁 sub-01/
    - 📁 anat/
      - 📄 sub-01\_T1w.nii.gz
    - 📁 func/
      - 📄 sub-01\_task-rest\_bold.nii.gz
      - 📄 sub-01\_task-rest\_bold.json
    - 📁 dwi/
      - 📄 sub-01\_dwi.nii.gz
      - 📄 sub-01\_dwi.json
      - 📄 sub-01\_dwi.bval
      - 📄 sub-01\_dwi.bvec
    - 📁 sub-02/
    - 📁 sub-03/
    - 📁 sub-04/

OPEN SCIENCE:

WHY



WHAT



HOW

# How to be Open – Choose your battles

## Be open when you can, as you can

### Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
<b>Citation standards</b>	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
<b>Data transparency</b>	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
<b>Analytic methods (code) transparency</b>	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
<b>Research materials transparency</b>	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
<b>Design and analysis transparency</b>	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
<b>Preregistration of studies</b>	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
<b>Preregistration of analysis plans</b>	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
<b>Replication</b>	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

# How to Open – You don't have to do it alone

- Training



- Asking for help

- Data Science and Sharing Team



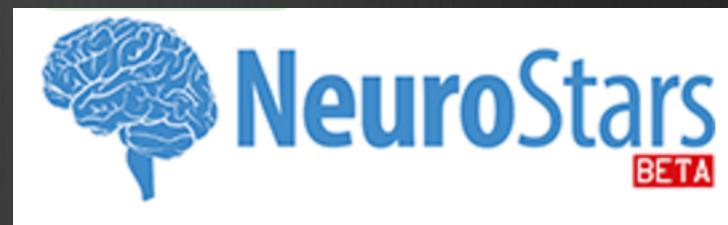
Adam Thomas



John Lee



Dylan Nielson



OPEN SCIENCE:

WHY

→

WHAT

→

HOW

# Data Science and Sharing Team's Workshop on Open and Reproducible Neuroscience

## Mar 13-17th, 2017

- 45 applications, 25 students attended
- 16 hours of instruction on Python, Git, Data Repositories, Biowulf integration, Pre-registration, and statistical rigor
- Instructors from Gallaudet, King's College London, AFNI and Biowulf Teams



- All course material available online:  
[https://github.com/nih-fmrf/NIMH\\_repro\\_2017](https://github.com/nih-fmrf/NIMH_repro_2017)
- Next course Nov 2017



# Summary and Take Homes

- Science is changing (for the better) in both scope (big) and culture (open) to address future challenges
- Open science strives to maximize reproducibility and transparency of data, code, and papers
- Adopting Open Science practices yields benefits in productivity, impact, and reach
- You don't have to do it all at once, and you don't have to do it alone

# Thanks!

See online slides for more URLs and references:

<https://github.com/agt24>

# Questions?