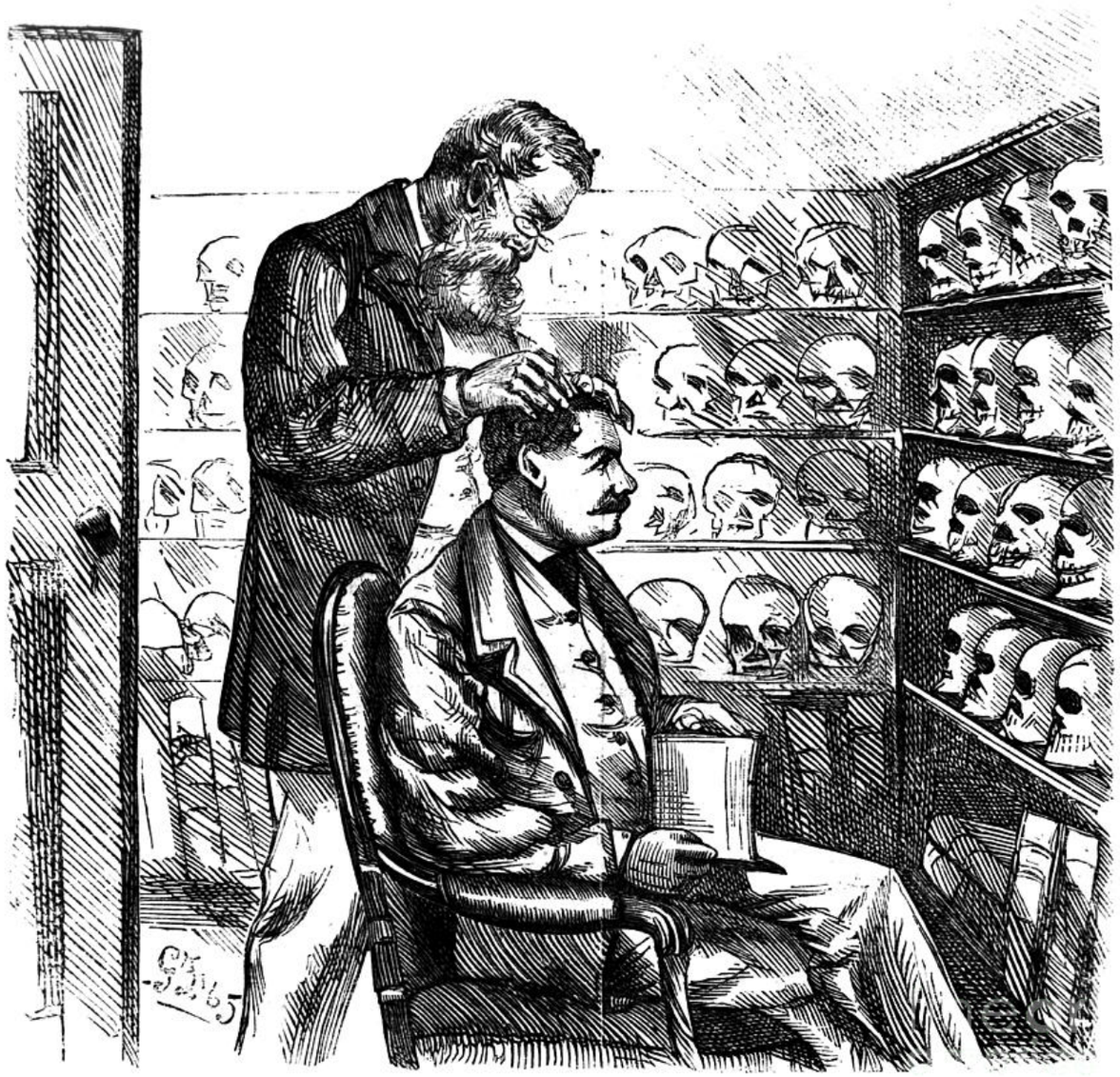
Stanford University

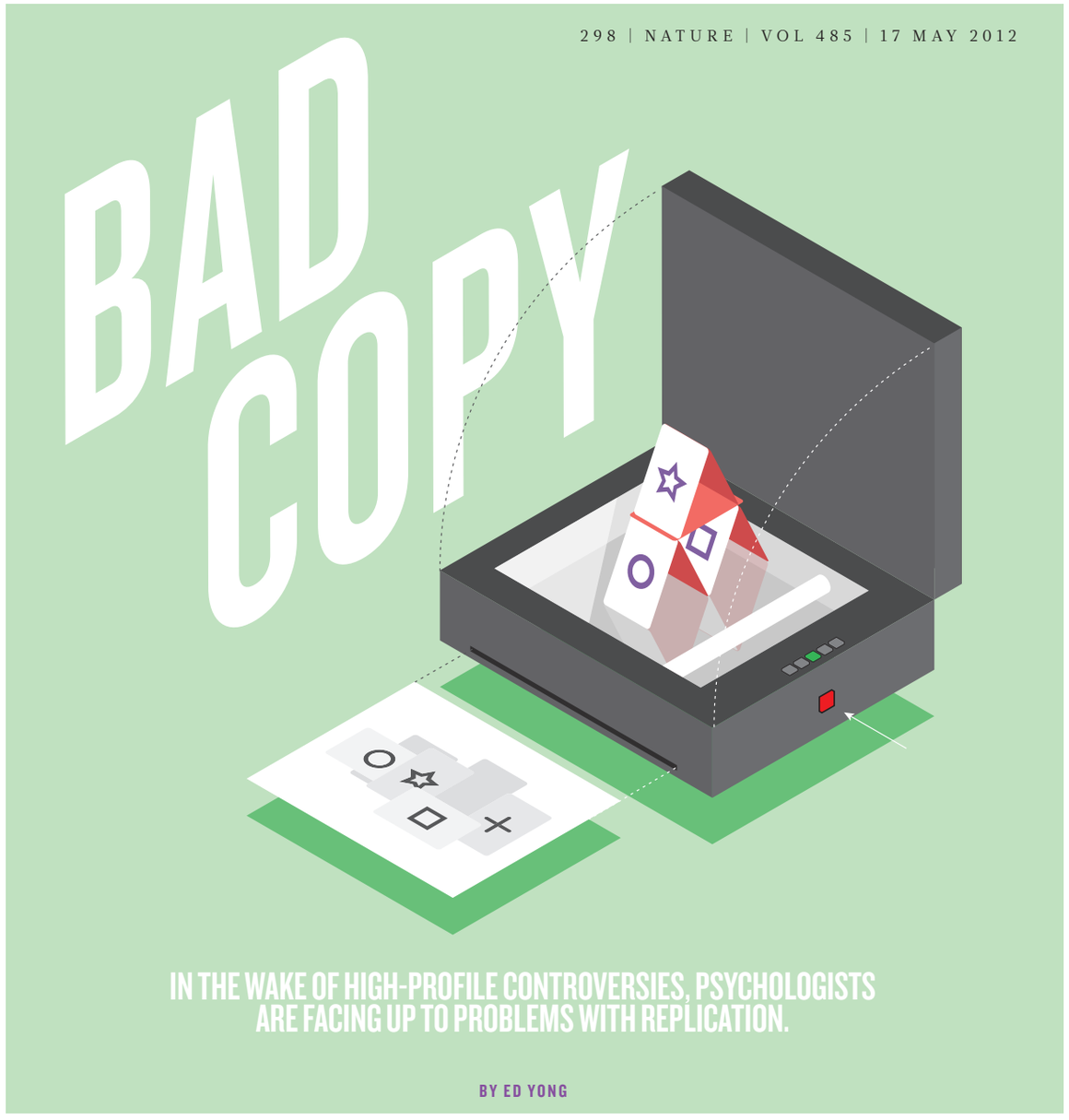
What's wrong with neuroimaging research, and how can we make it right?

Russell Poldrack
Stanford University



A HINT TO PHRENOLOGISTS; or, "September 20, 1878."

Science in crisis (?)



Rigorous replication effort succeeds for just two of five cancer papers

By **Jocelyn Kaiser** | Jan. 18, 2017 , 1:00 PM

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

29 MARCH 2012 | VOL 483 | NATURE | 531

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

SCIENCE sciencemag.org 28 AUGUST 2015 • VOL 349 ISSUE 6251

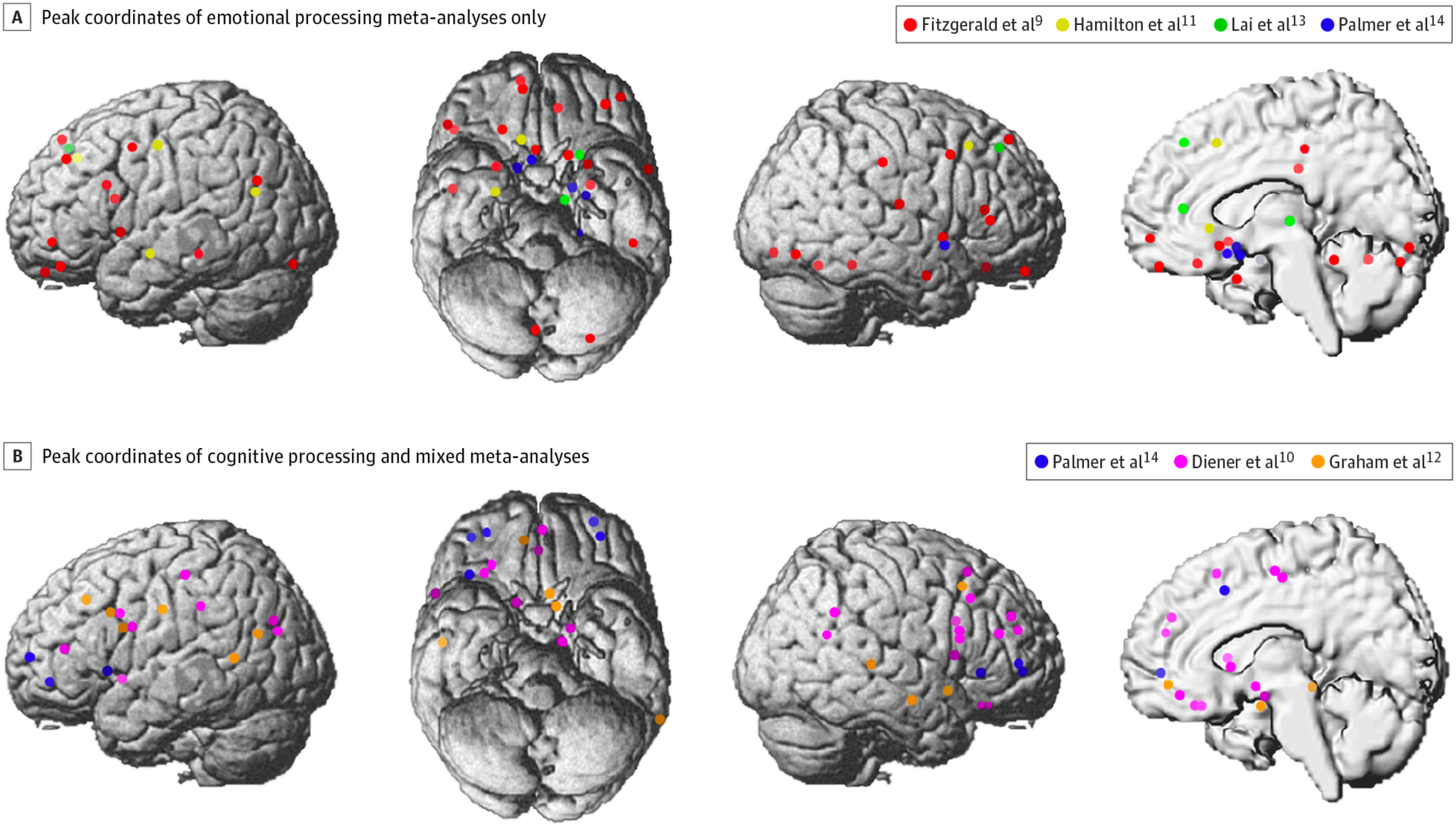
We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available.

Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results

A full-scale replication project for neuroimaging would be far too expensive

How can we know whether our results are reliable?

What is the brain dysfunction in major depression?
Meta-analysis of 99 published studies



Muller et al., 2017, *JAMA Psychiatry*

Altered Brain Activity in Unipolar Depression Revisited

Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD[1,2]; Edna C. Cieslik, PhD[1,2]; Ilinca Serbanescu, MSc[1]; et al

JAMA Psychiatry. 2017;74(1):47-55.

Findings This conceptual replication of meta-analyses of 99 neuroimaging experiments in unipolar depression did not reveal any convergence, which is at odds with the findings of previous meta-analyses.

Meta-analyses Across Emotional Experiments

None of the 9 emotional meta-analyses revealed any significant results (all emotional: 65 experiments [$P > .69$]; in-

Meta-analyses Across Cognitive Experiments

None of the 4 cognitive meta-analyses revealed any significant results (all cognitive: 34 experiments [$P > .63$]; in-

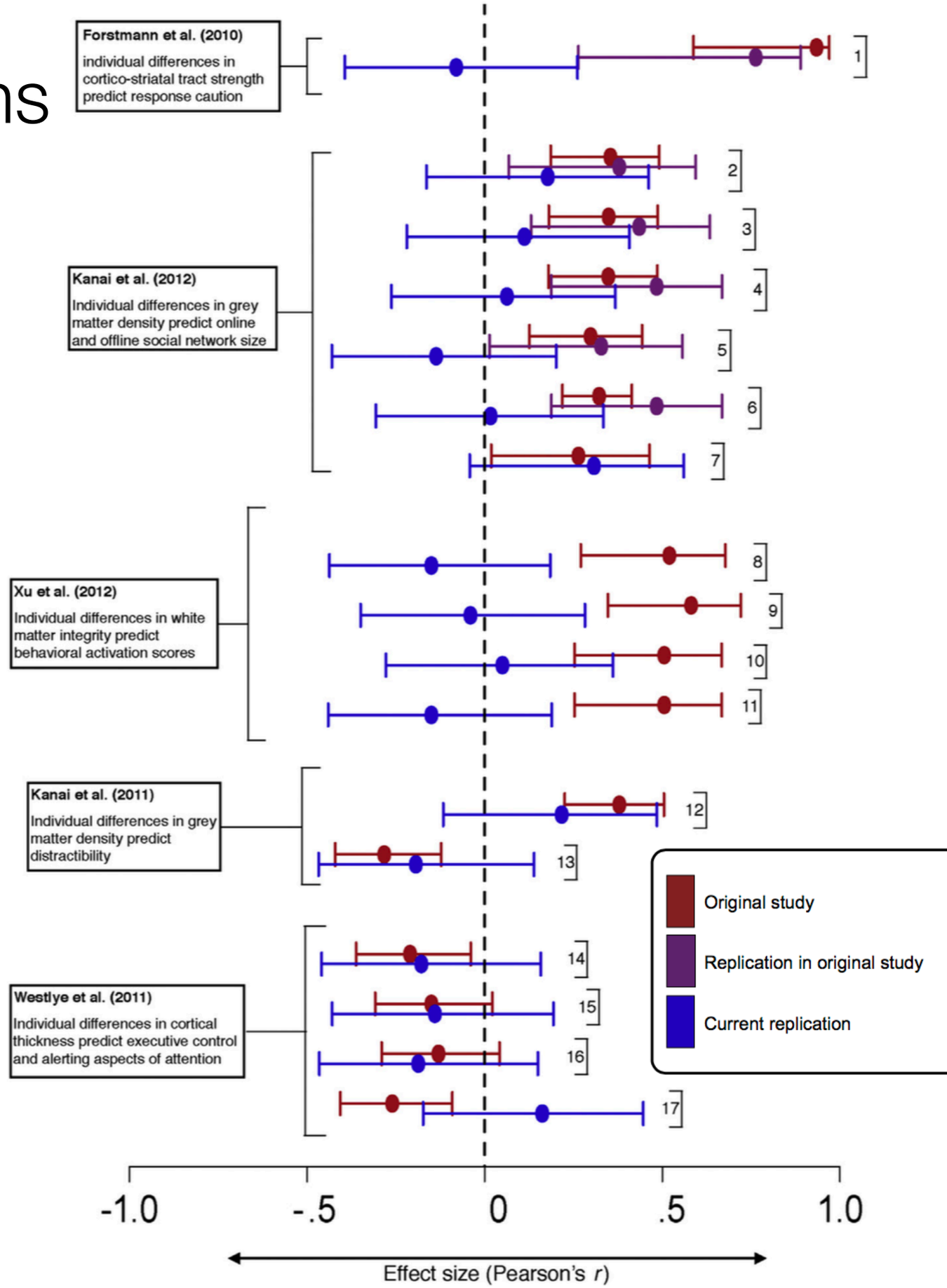
Irreproducible correlations

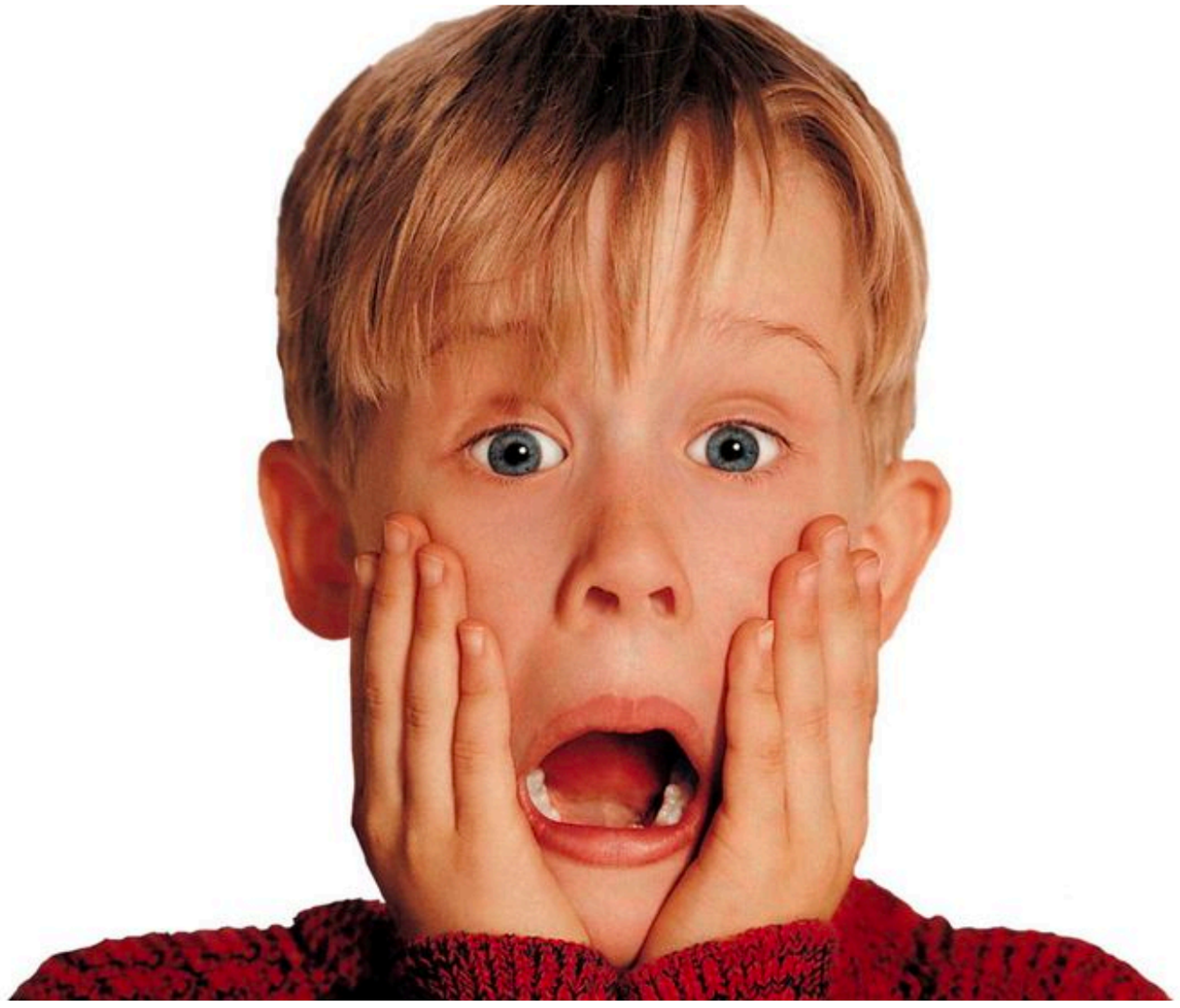
A purely confirmatory replication study of structural brain-behavior correlations

Wouter Boekel [a,], Eric-Jan Wagenmakers [a], Luam Belay [a], Josine Verhagen [a], Scott Brown [b] and Birte U. Forstmann [a]*

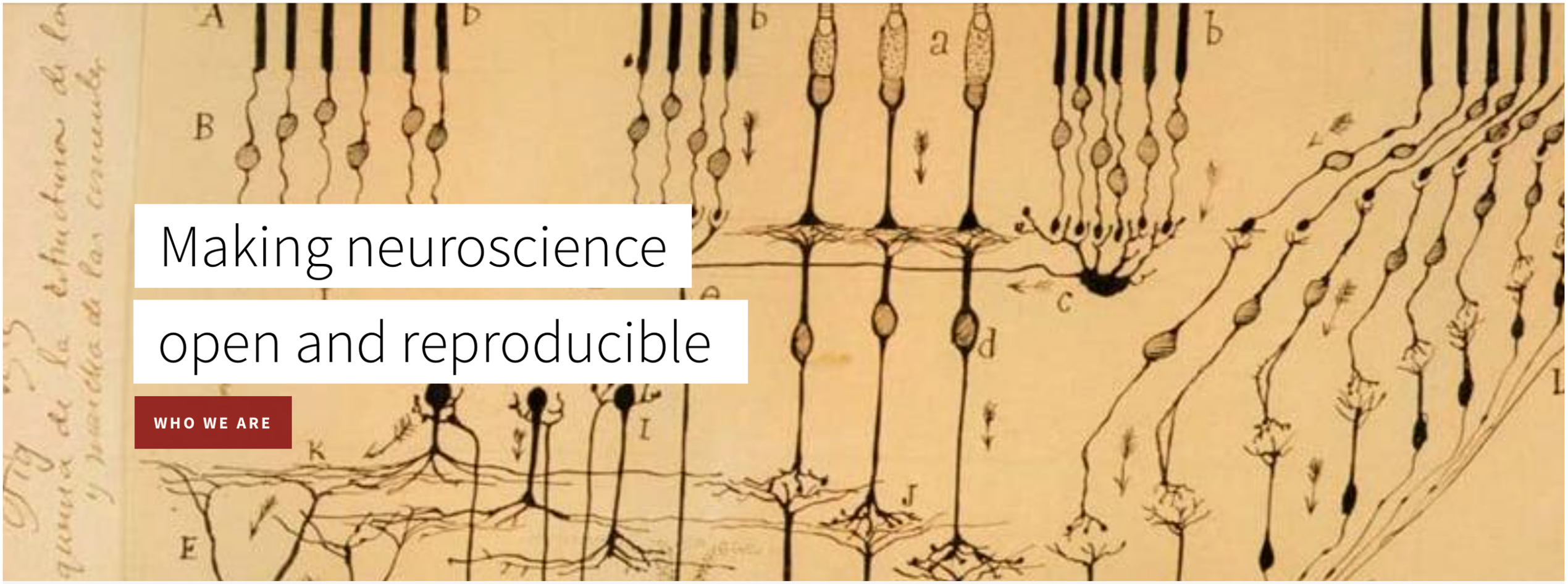
CORTEX 66 (2015) 115–133

Here, we attempt to replicate five structural brain-behavior correlation studies comprising a total of 17 effects. To prevent the impact of QRPs we employed a preregistered, purely confirmatory replication approach. For all but one of the 17 findings under scrutiny, confirmatory Bayesian hypothesis tests indicated evidence in favor of the null hypothesis ranging from anecdotal (Bayes factor < 3) to strong (Bayes factor > 10). In several studies, effect size estimates were substantially lower than in the original studies.





We seem to have created quite a mess.
How can we fix it?



Making neuroscience open and reproducible

WHO WE ARE

Reproducibility matters

Neuroscience research is the basis for critical decisions about health and society. Our first goal as researchers is to ensure that the results of our research will stand the test of time.

Enabling better research

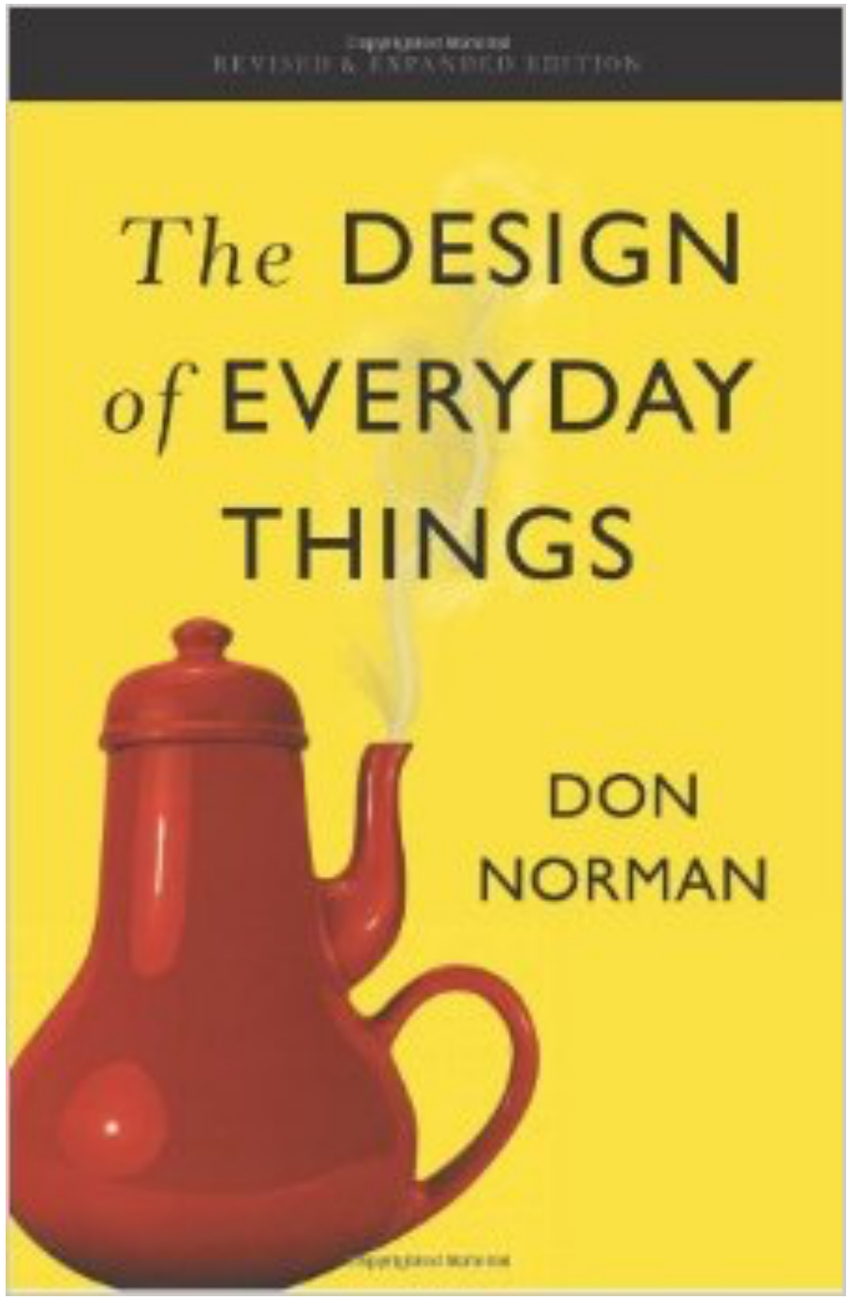
We are expanding the OpenfMRI project into a free and open platform that will enable the analysis and sharing of neuroimaging data, harnessing the power of high-performance computing to improve the quality of research.

From data to discovery

Our platform will provide neuroimaging researchers with leading-edge tools to analyze and share large datasets, with a focus on quantifying the reproducibility of the results.

http://reproducibility.stanford.edu

Designing a more reproducible scientific enterprise

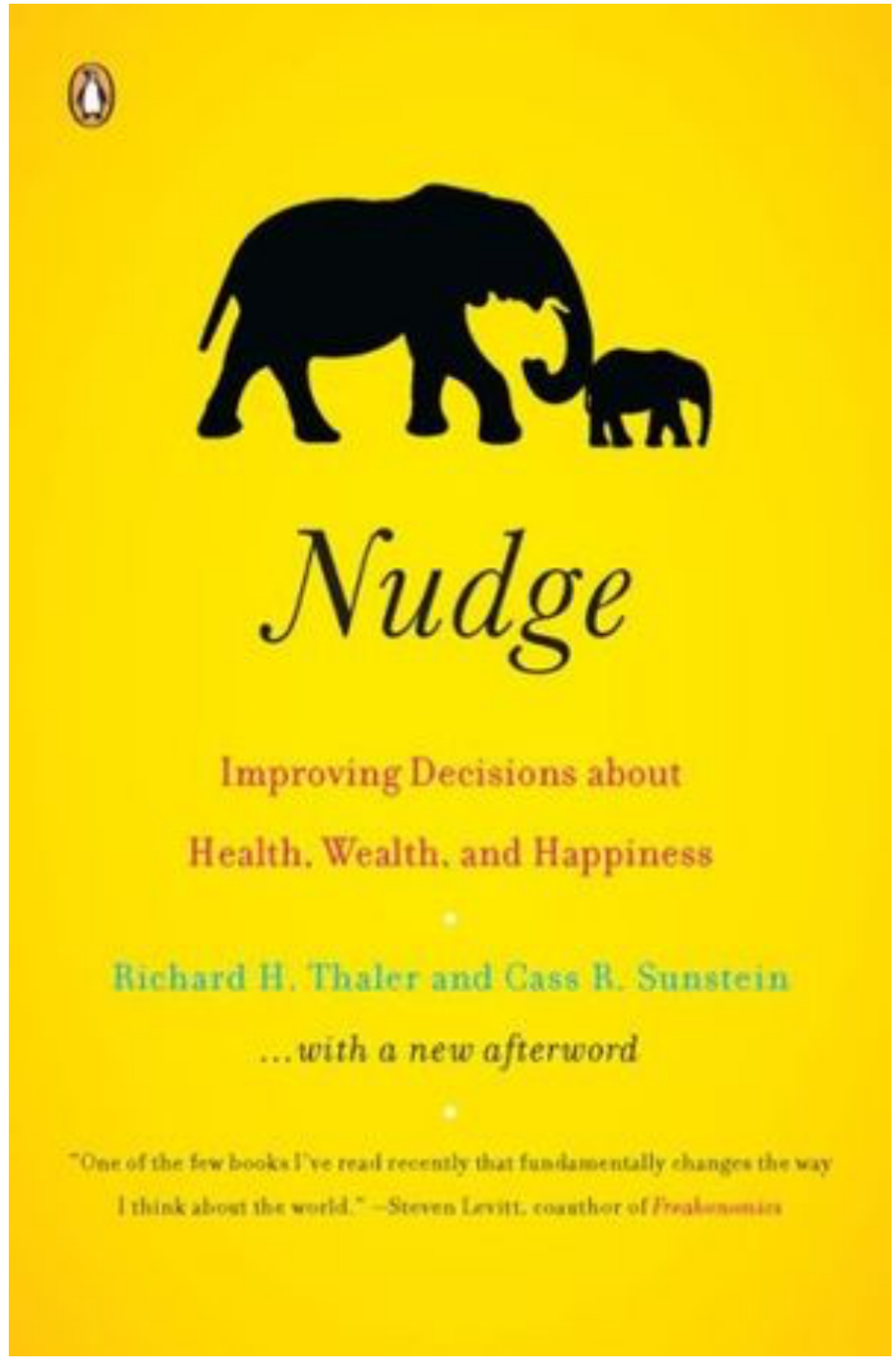


FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

Improving the choice architecture of science

- Choice architecture
  - particular set of features that drive people toward or away from particular choices
- Nudges
  - Improving incentives
  - Using the power of defaults
  - Providing feedback
  - Expecting and prevent errors



Stanford University

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

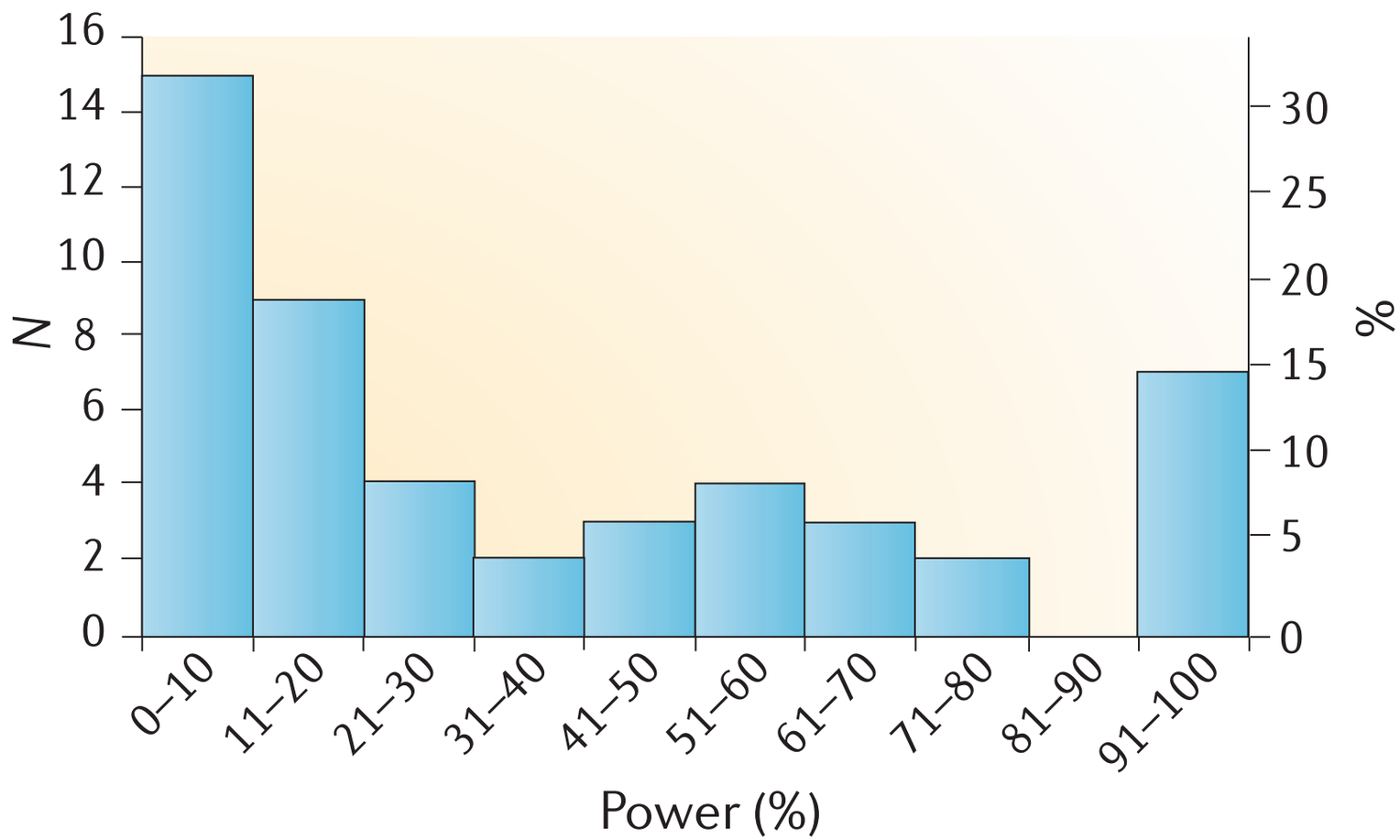
- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.

Neuroscience research is badly underpowered

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

NATURE REVIEWS | NEUROSCIENCE VOLUME 14 | MAY 2013 | 365

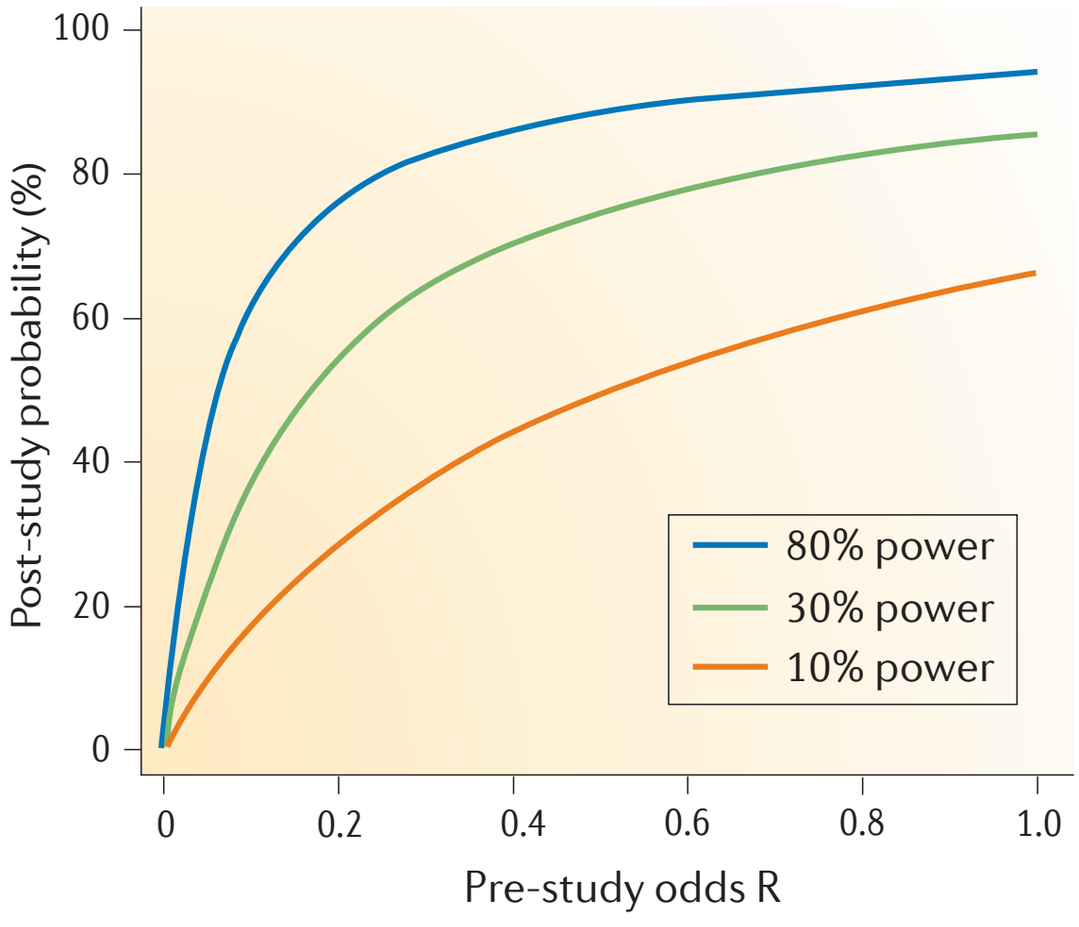


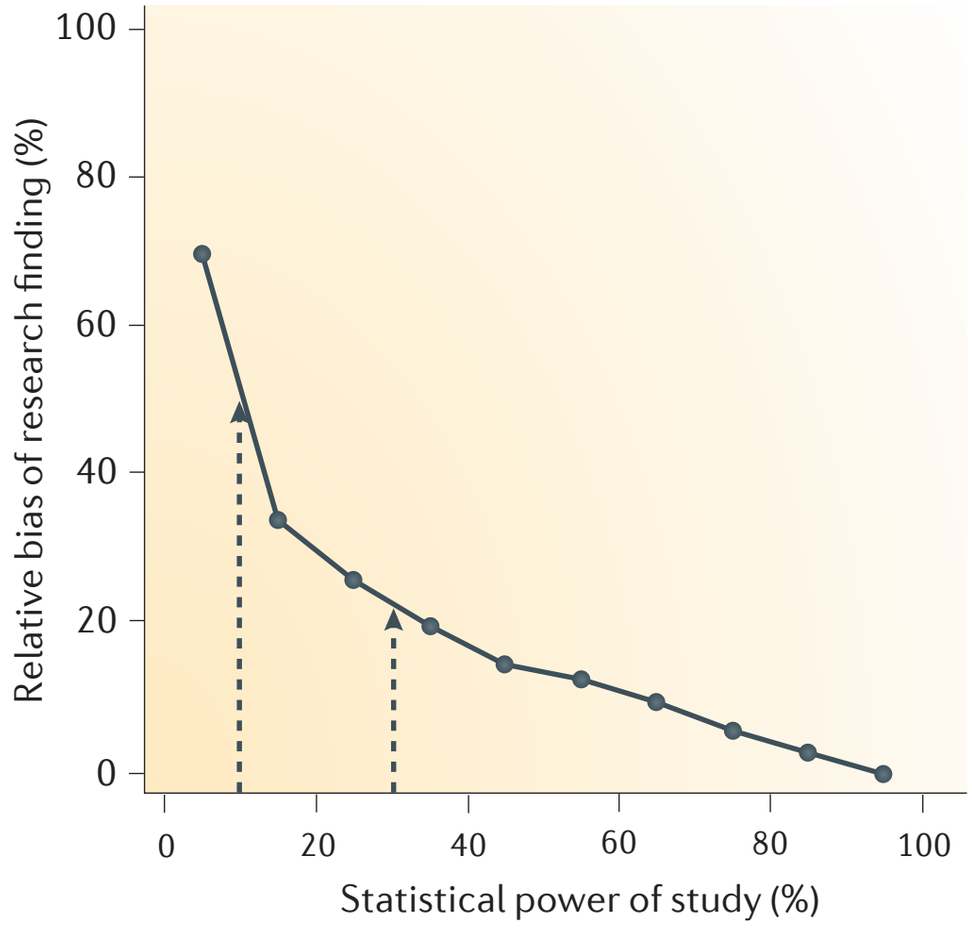
Low power -> unreliable science

Positive Predictive Value (PPV): The probability that a positive result is true

Winner's Curse: overestimation of effect sizes for significant results

$$PPV = ([1 - \beta] \times R) / ([1 - \beta] \times R + \alpha)$$



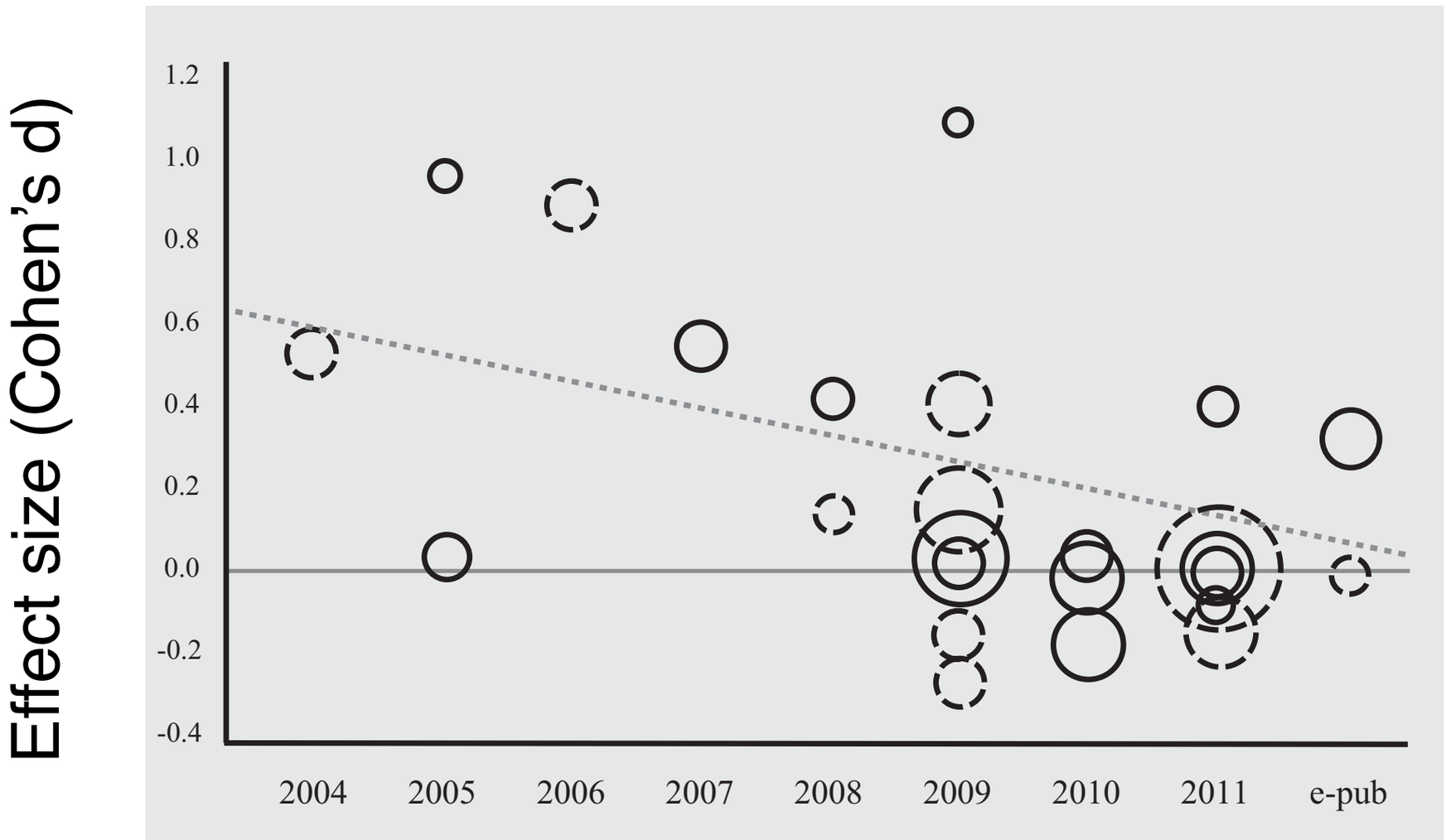


Button et al., 2013

Small samples + publication bias: the case of candidate genes

A Systematic Review and Meta-Analysis on the Association Between BDNF val[66]met and Hippocampal Volume—A Genuine Effect or a Winners Curse?

Marc L. Molendijk,[1,2]* Boudewijn A.A. Bus,[3] Philip Spinhoven,[1,2,4] Anna Kaimatzoglou,[1] Richard C. Oude Voshaar,[5] Brenda W.J.H. Penninx,[4,5,6] Marinus H. van IJzendoorn,[7,8] and Bernet M. Elzinga[1,2]



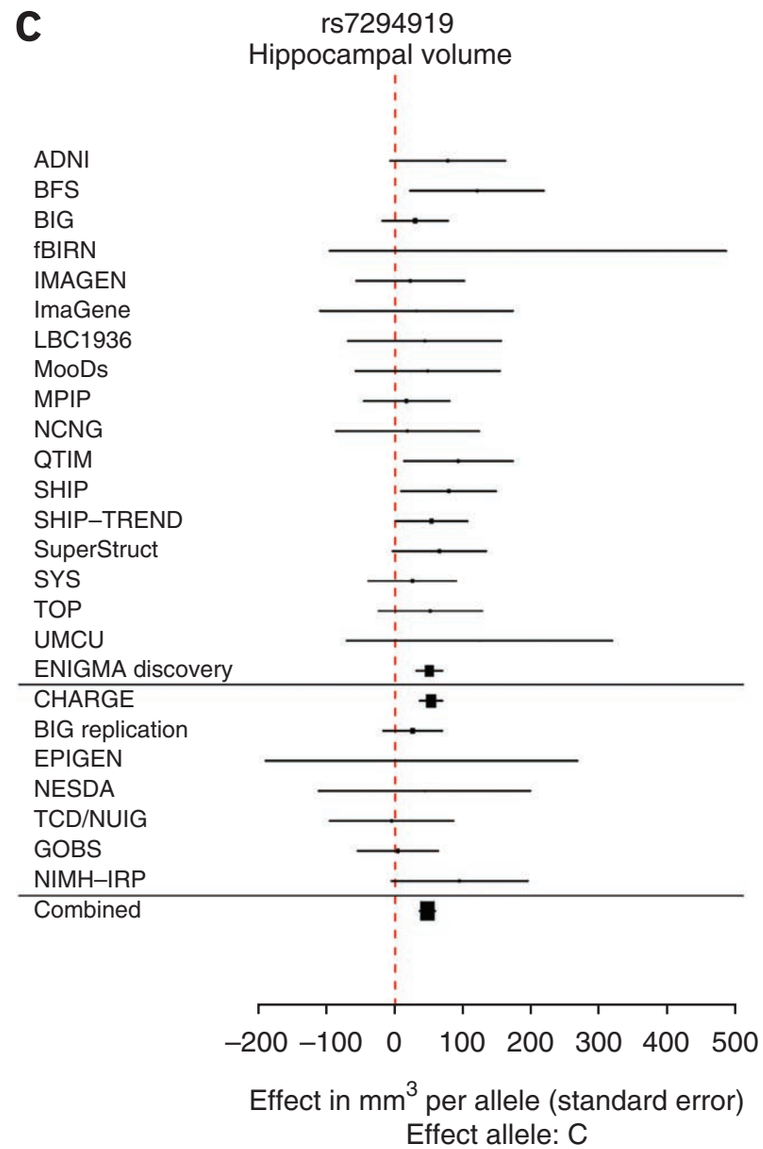
Candidate gene associations fail in well-powered GWAS

Identification of common variants associated with human hippocampal and intracranial volumes

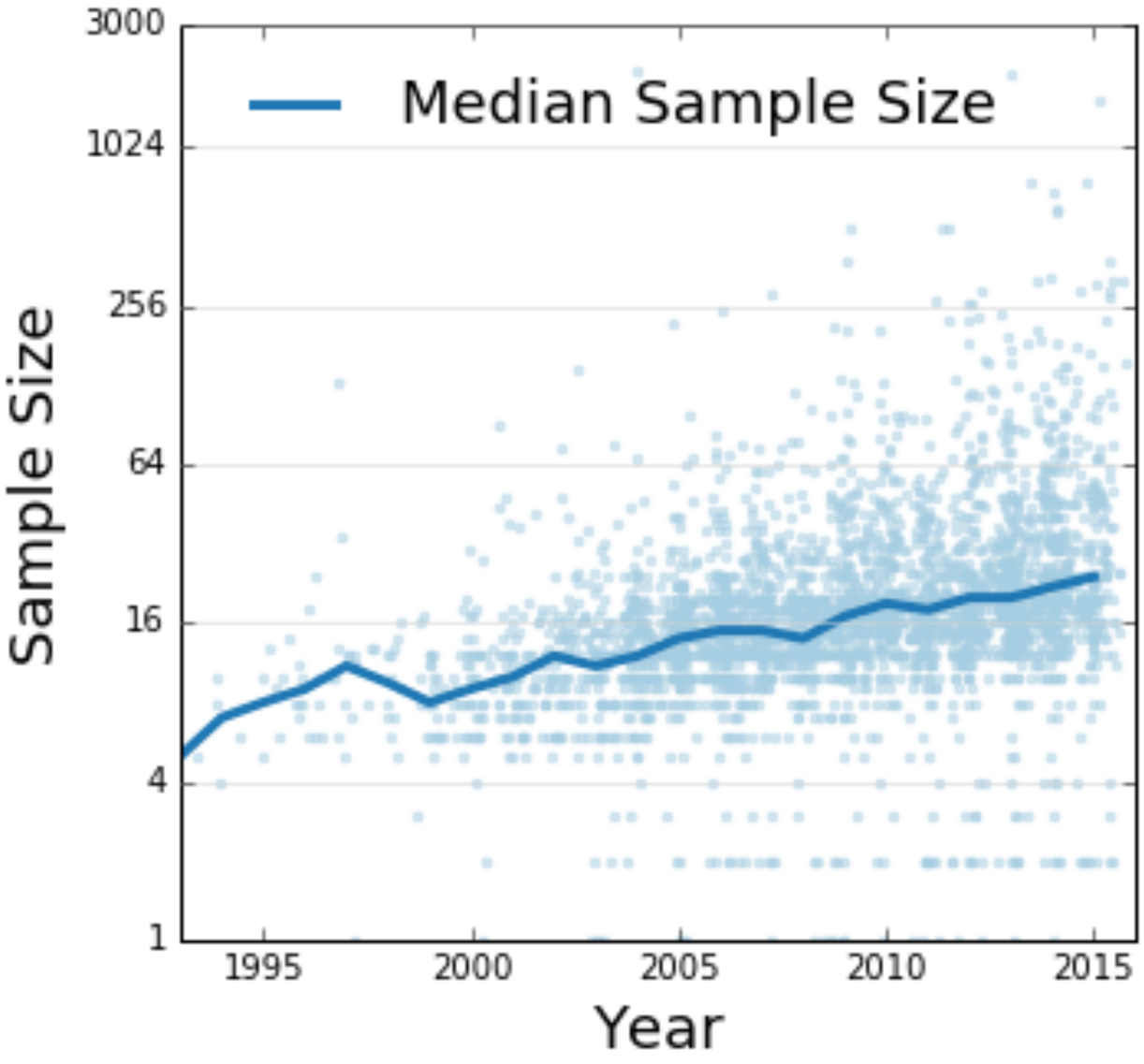
Jason Stein et al. for the Enigma Consortium

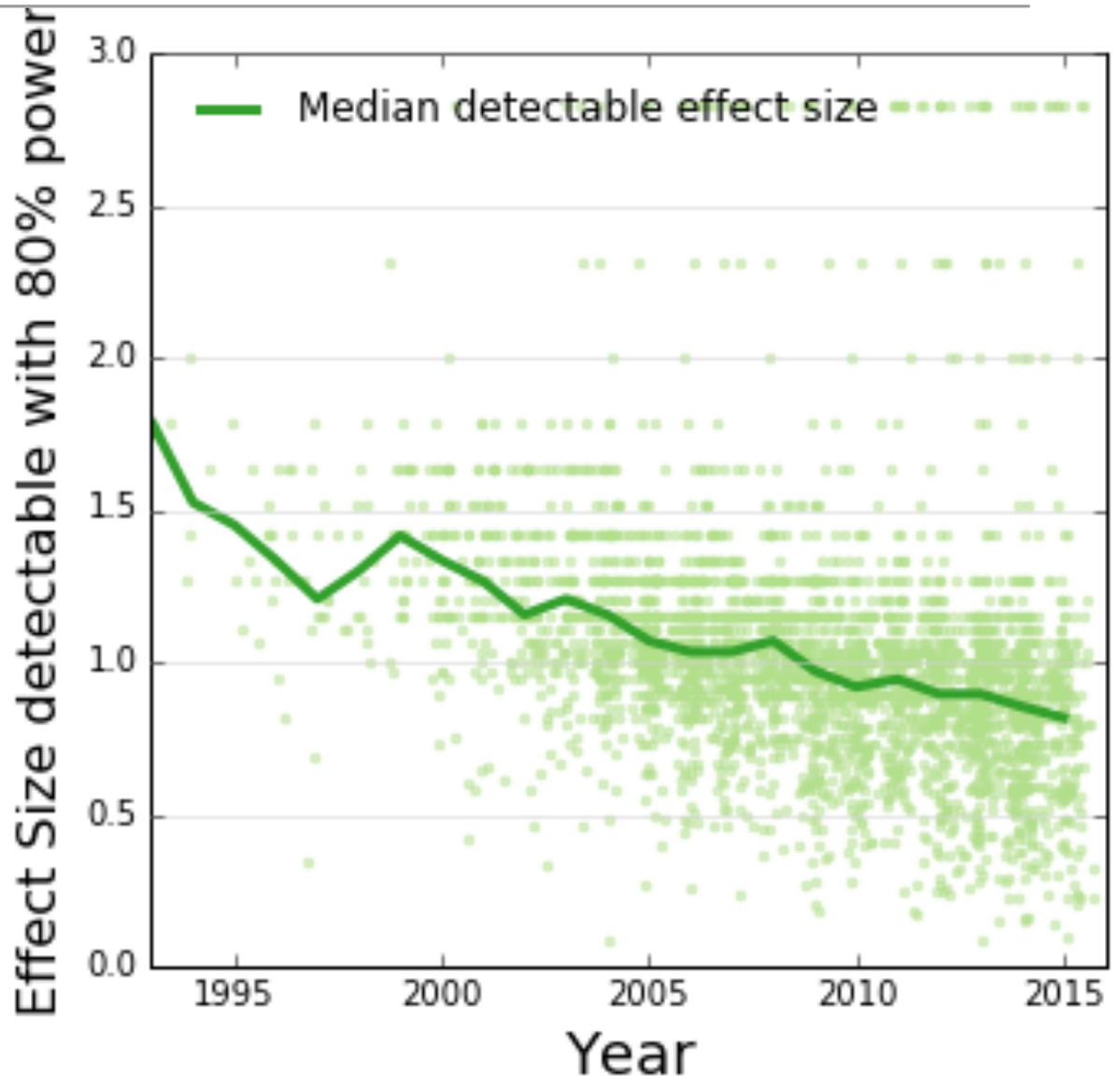
VOLUME 44 | NUMBER 5 | MAY 2012 **NATURE GENETICS**

In general, previously identified polymorphisms associated with hippocampal volume showed little association in our meta-analysis (*BDNF*, *TOMM40*, *CLU*, *PICALM*, *ZNF804A*, *COMT*, *DISC1*, *NRG1*, *DTNBP1*), nor did SNPs previously associated with schizophrenia or bipolar disorder



How well powered are fMRI studies?



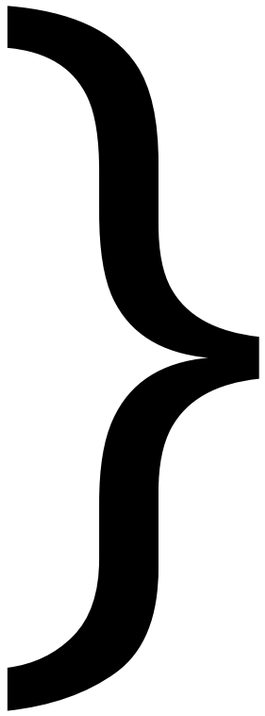


- Median study in 2015 was powered for find a single 200 voxel activation with d~0.75
- Is that plausible?

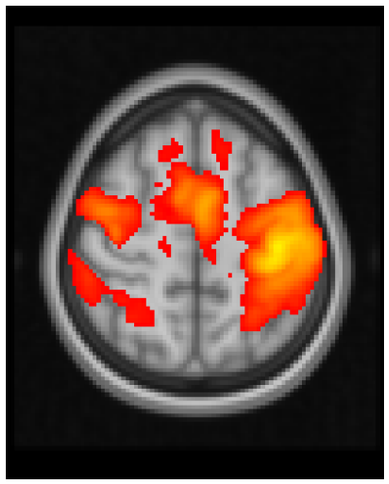
Poldrack et al, 2017, NRN

Estimating realistic effect sizes

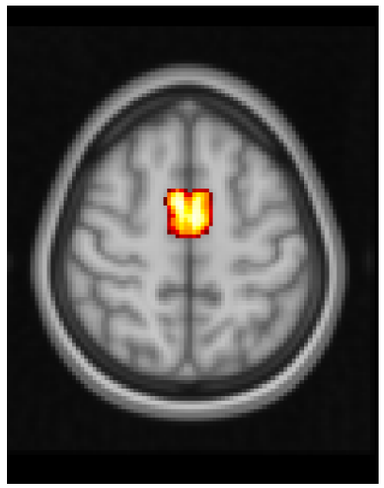
HCP motor task data



group analysis



group map



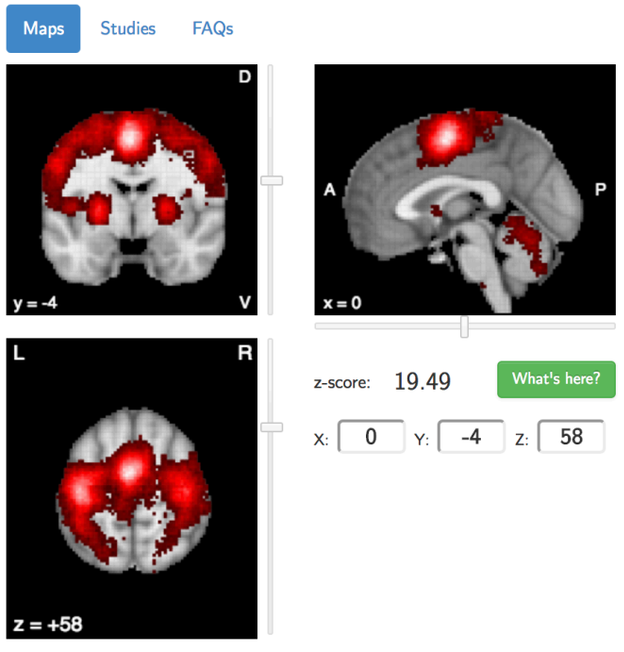
Regions of Interest

Primary Motor; Premotor; Striatum; Retinotopic Visual Areas

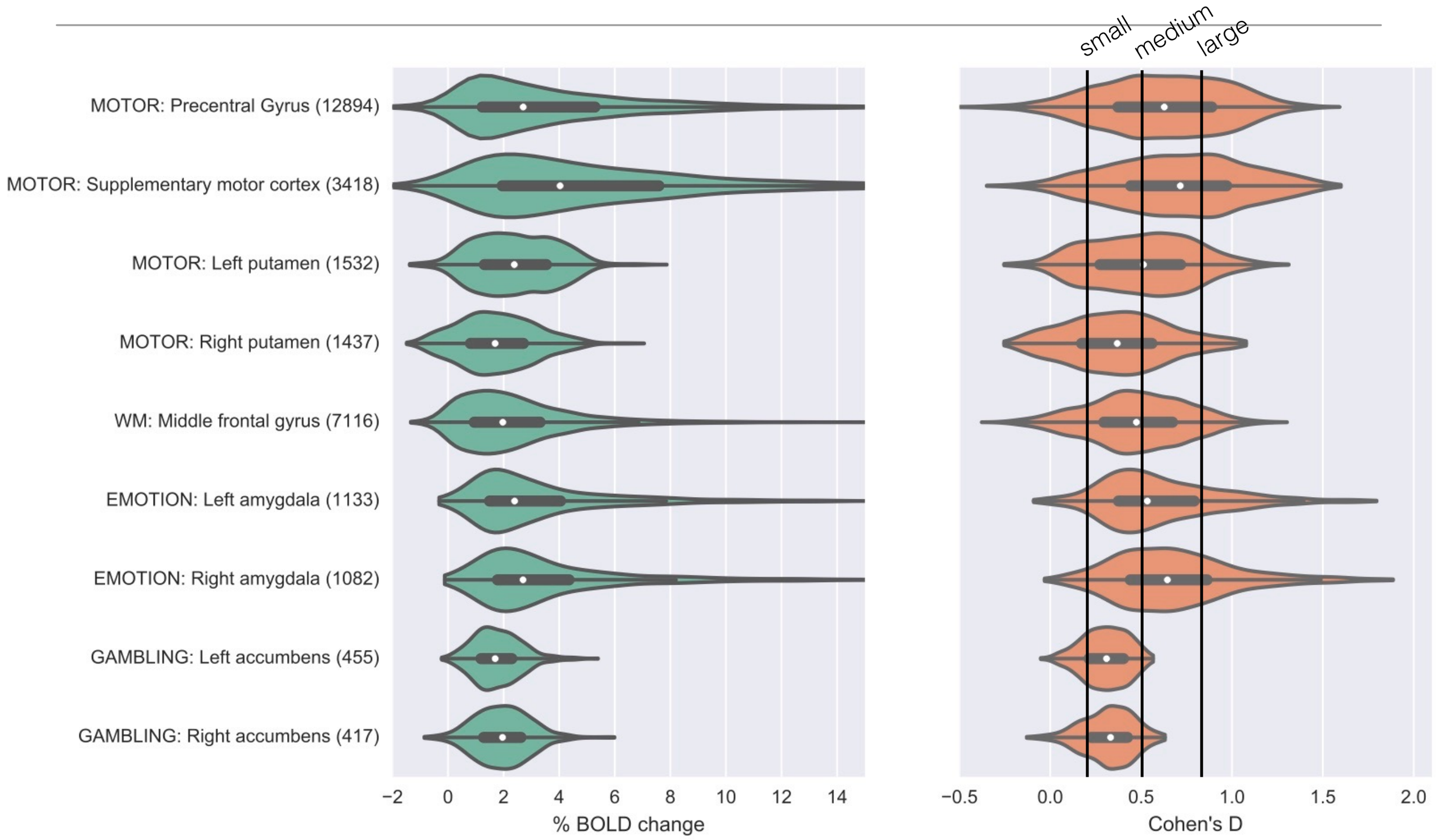
motor

unbiased ES estimate

An automated meta-analysis of 2081 studies



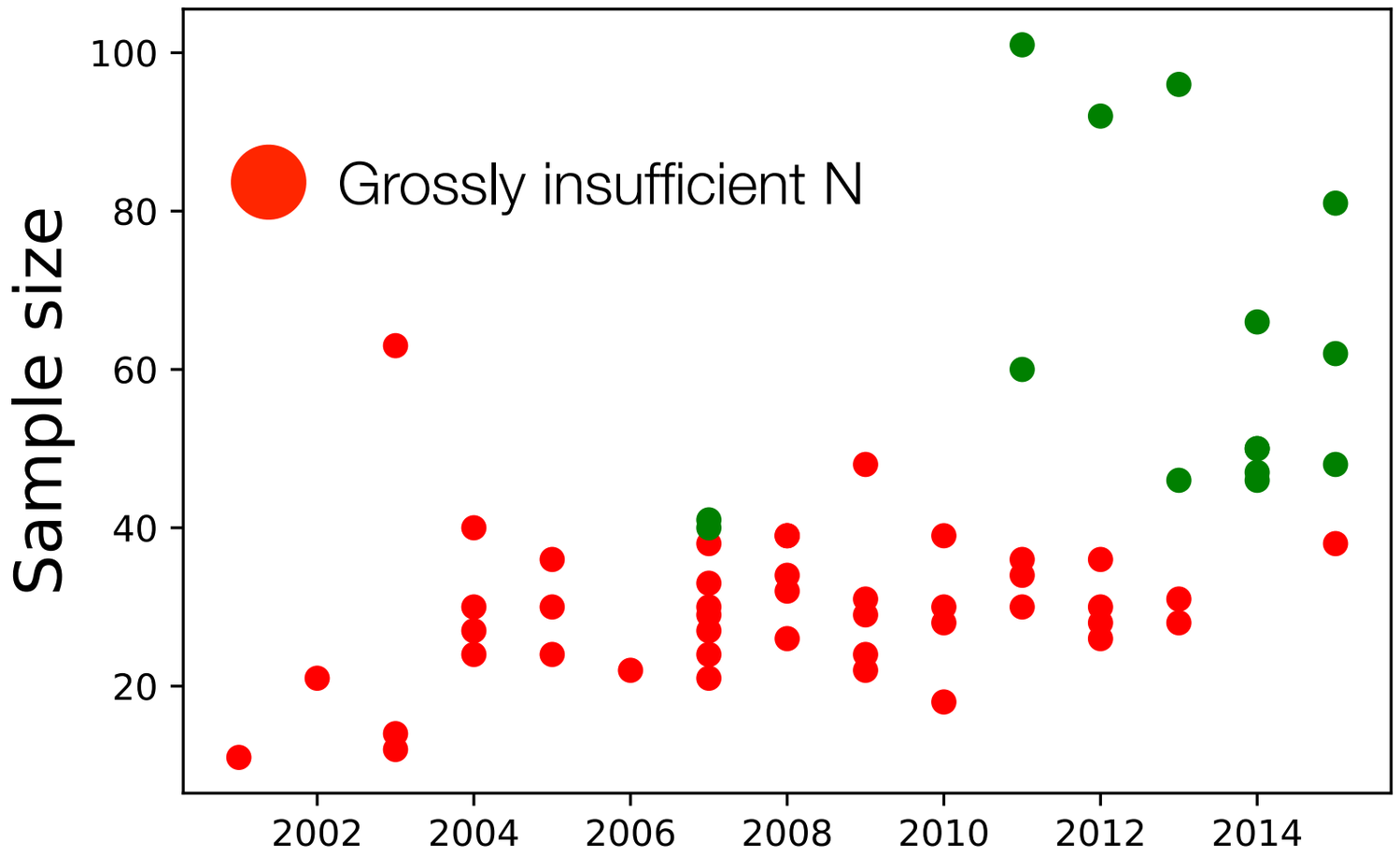
What are realistic effect sizes for fMRI?



Estimated from HCP task data
using combined anatomical + neurosynth ROIs

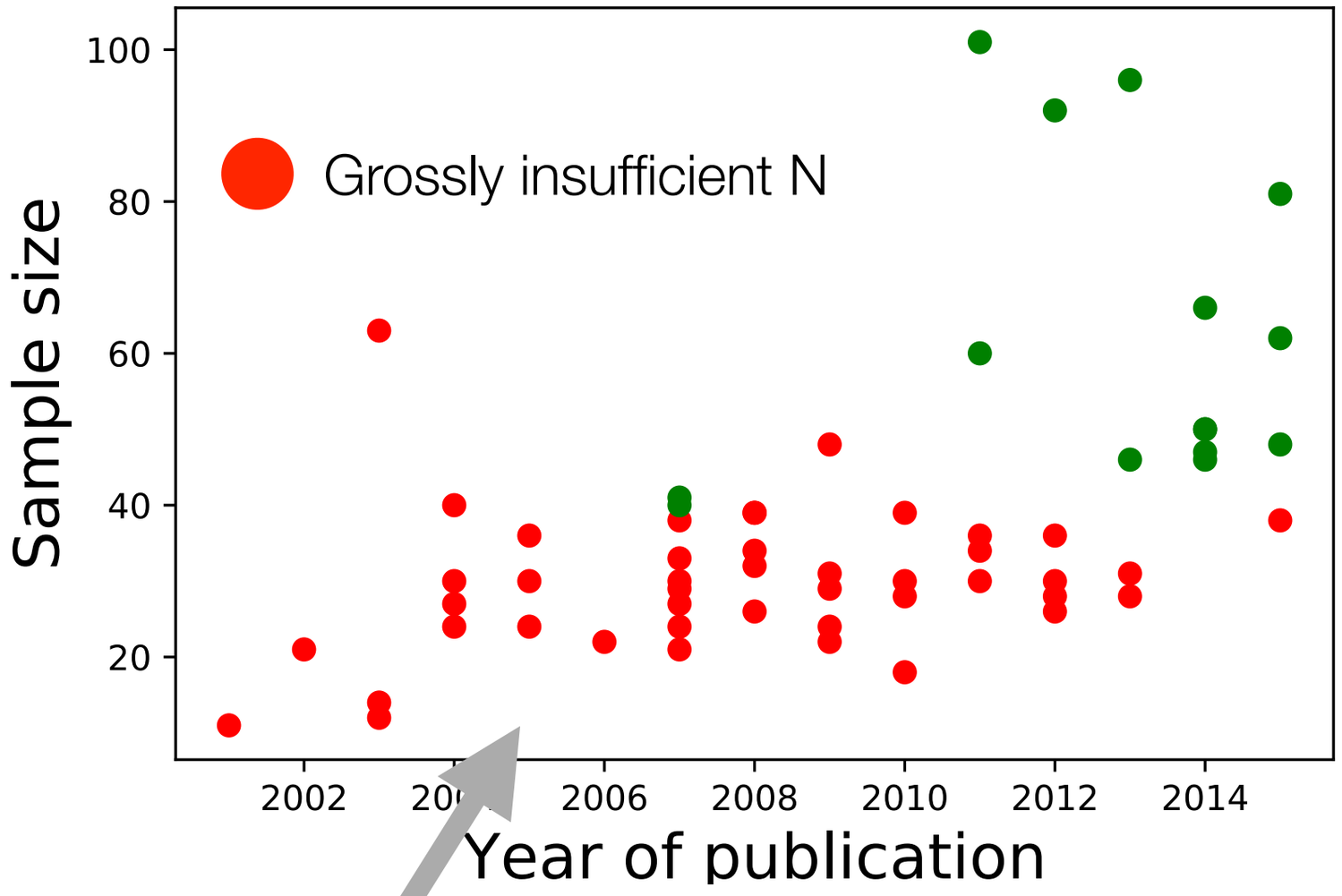
Poldrack et al, 2017, NRN

Depression studies from Muller et al.



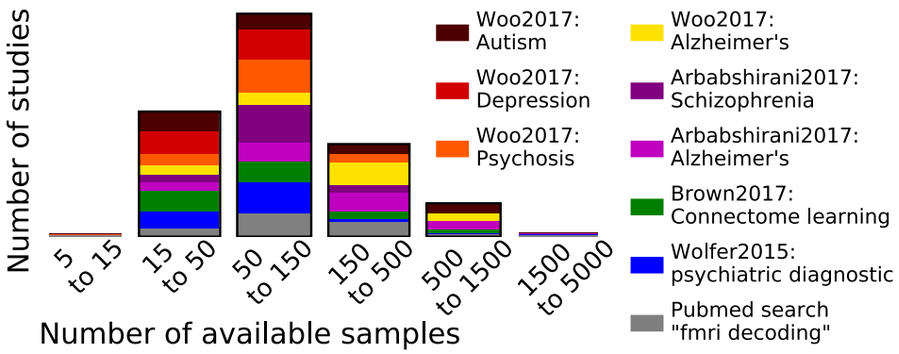
Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification. This requirement offers extra protection for the first requirement. *Samples smaller than 20 per cell are simply not powerful enough to detect most effects, and so there is usually no good reason to decide in advance to collect such a small number of observations.* Smaller samples, it follows, are much more likely to reflect interim data analysis and a flexible termination rule (Simmons et al., 2011)

Depression studies from Muller et al.

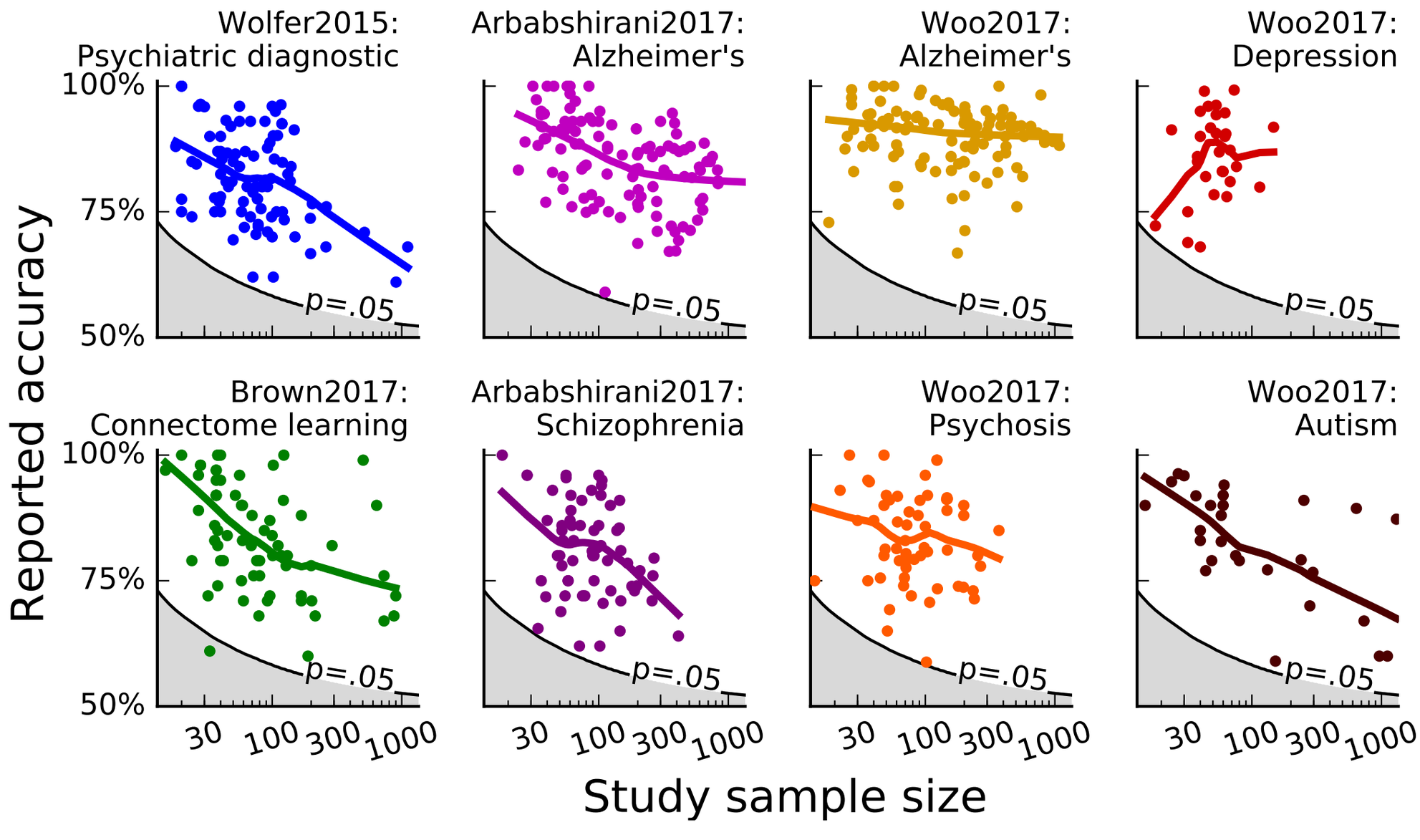


634 individuals diagnosed with depression volunteered for fMRI studies that were were almost certain to generate false results due to insufficient power

Small samples inflate predictive accuracy estimates



Varoquaux, 2017



Doing well-powered science as an early-career researcher

Neuron
NeuroView

The Costs of Reproducibility

Russell A. Poldrack[1,*]

Neuron *101*, January 2, 2019

- Underpowered science is futile, but many ECRs don't have resources to do sufficiently powered studies
- "if you can't answer the question you love, love the question you can" (Kanwisher, 2017)
- Pivots:
  - Collaborate
  - Use shared data
  - Focus on theory/computational modeling

Data sharing success story

- Mac Shine (Univ. of Sydney)
  - Poldracklab postdoc, 2014-2016
  - Collected no new data
  - Published papers in *PNAS*, *Neuron*, *J. Neuroscience*, *Nature Neuroscience*, and *Network Neuroscience*
- all based on shared data!

Stanford University

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.

Threats to reproducibility: Methodological flexibility

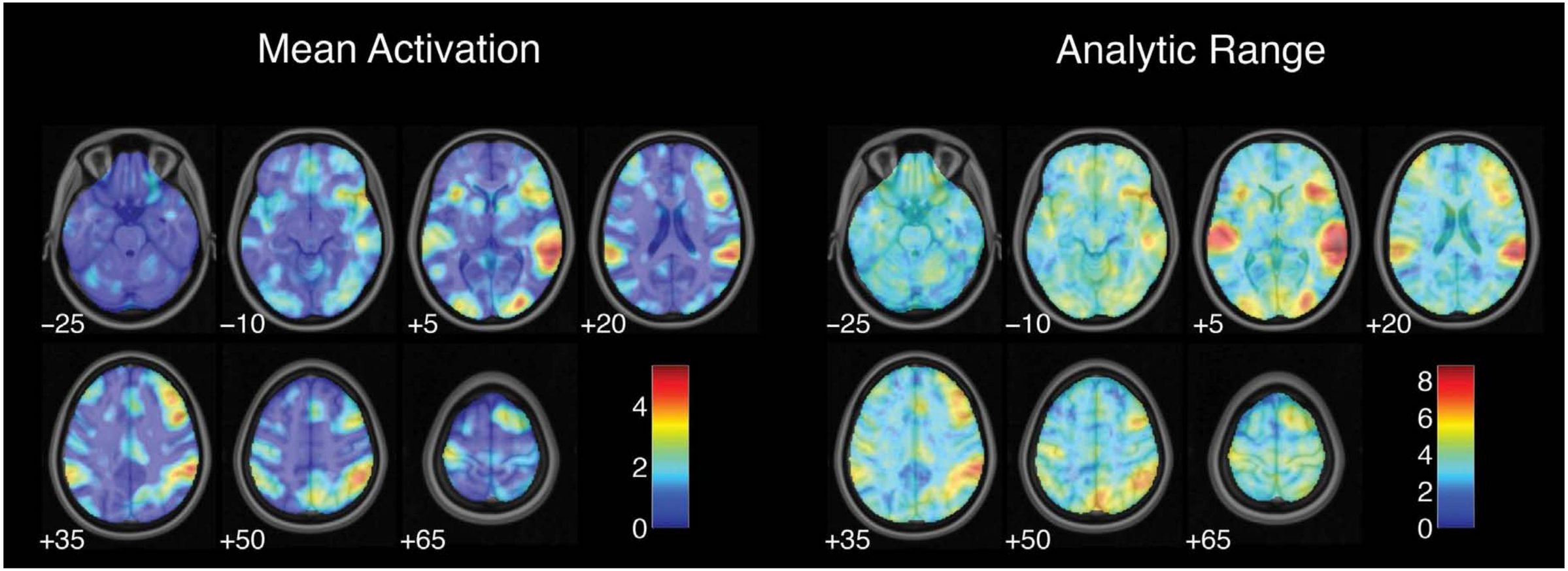
frontiers in NEUROSCIENCE

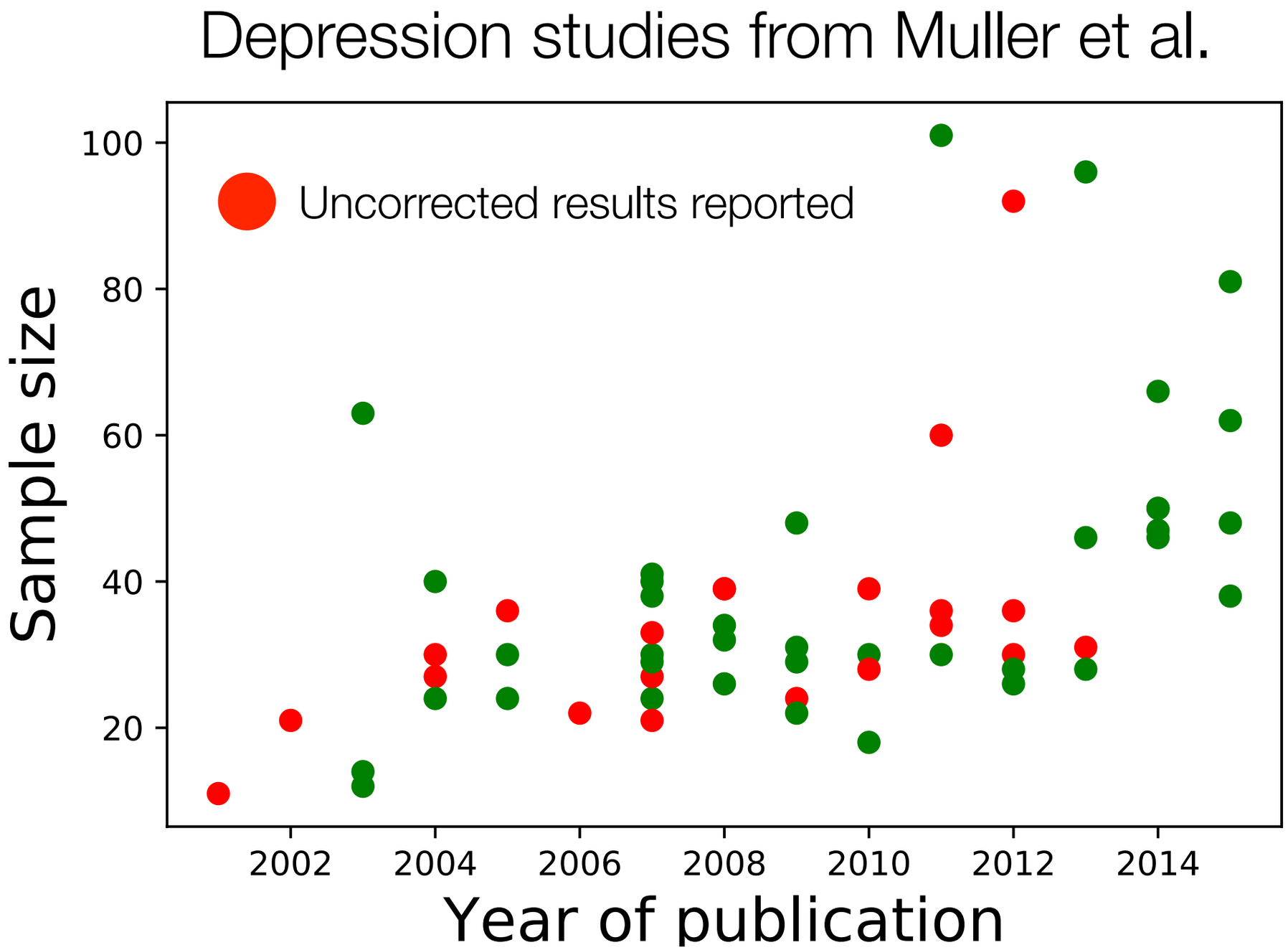
ORIGINAL RESEARCH ARTICLE
published: 11 October 2012
doi: 10.3389/fnins.2012.00149

On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments

Joshua Carp*

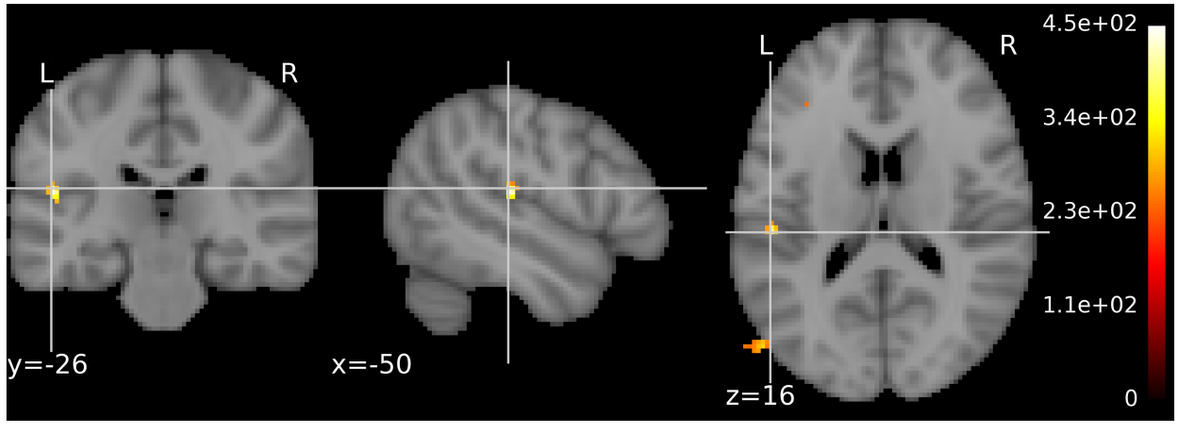
6,912 pipelines



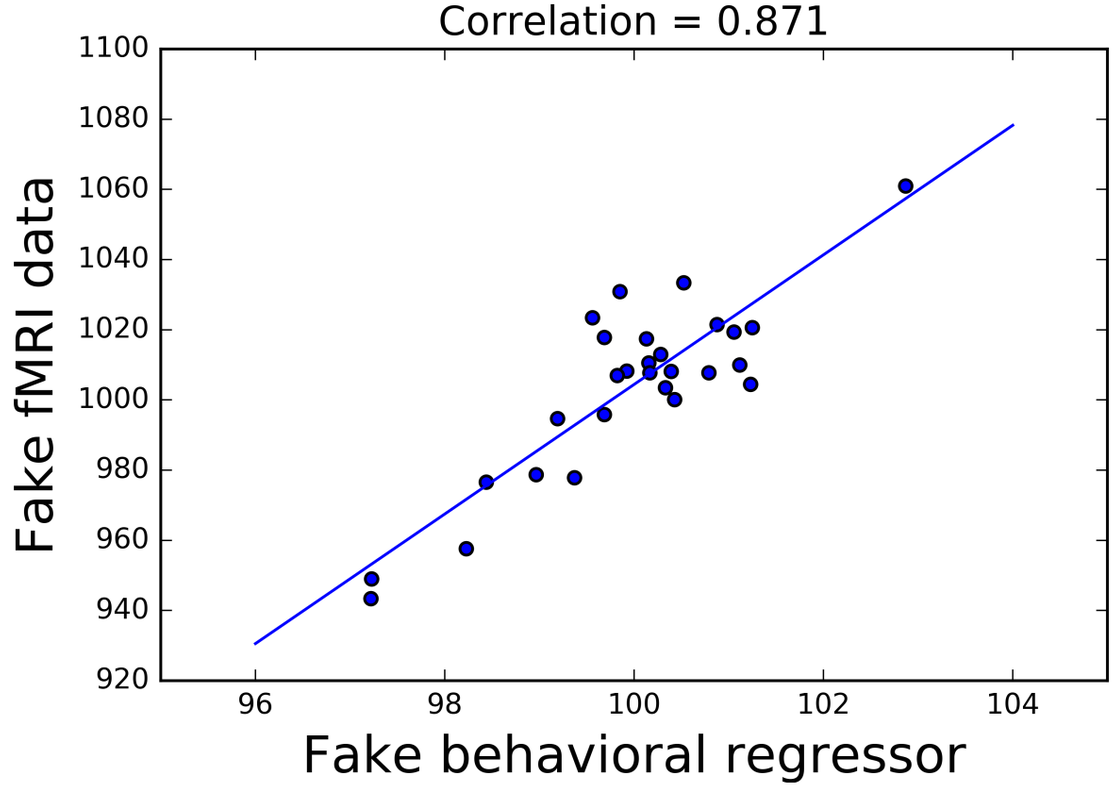

Depression studies from Muller et al.
Uncorrected results reported
Sample size
100
80
60
40
20
2002
2004
2006
2008
2010
2012
2014
Year of publication

Without correction, random data can yield beautiful results!

Correlation between random simulated behavioral variable and activation across 28 subjects



~220,000 voxels
$p < 0.001$
10 voxels cluster threshold



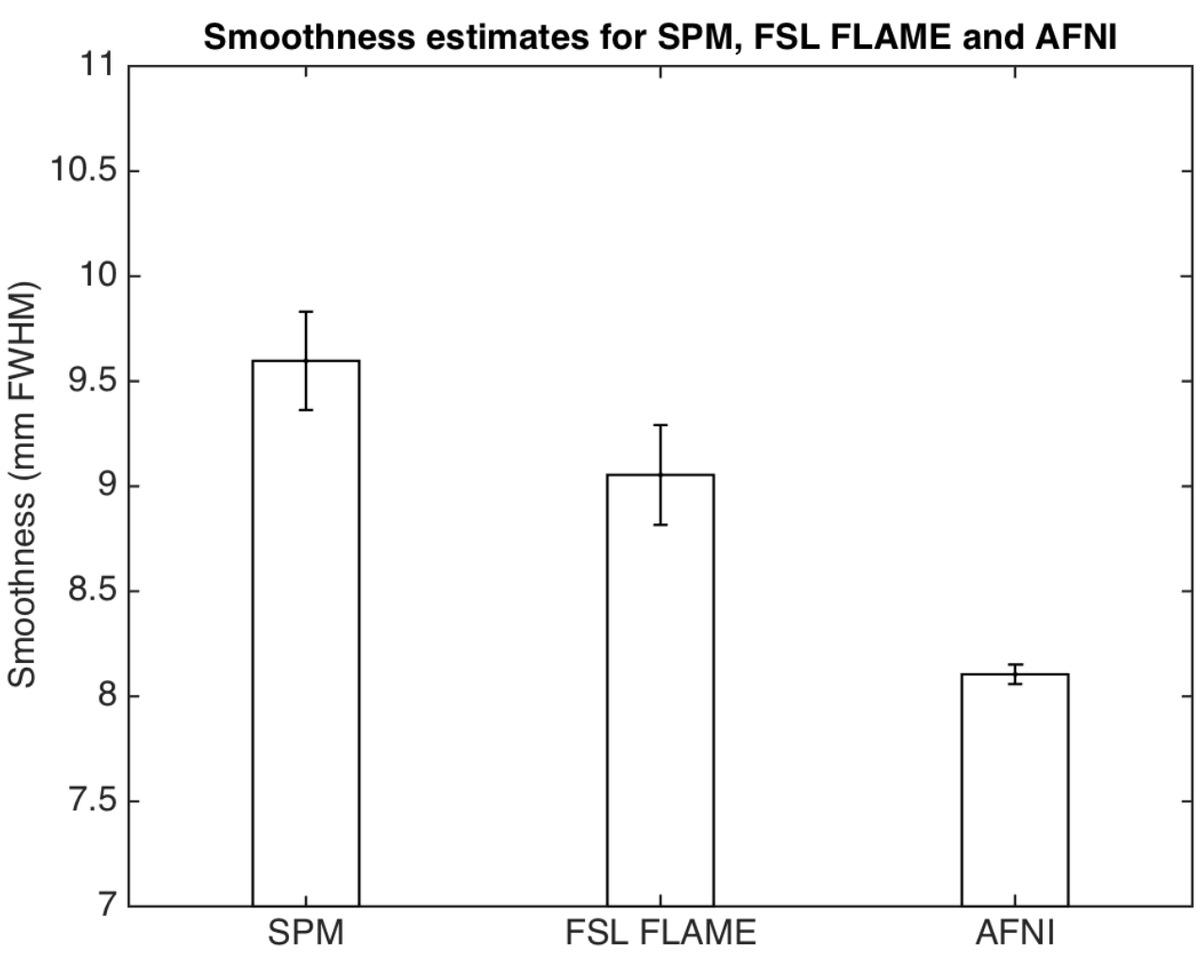
https://github.com/poldracklab/ScanningTheHorizon

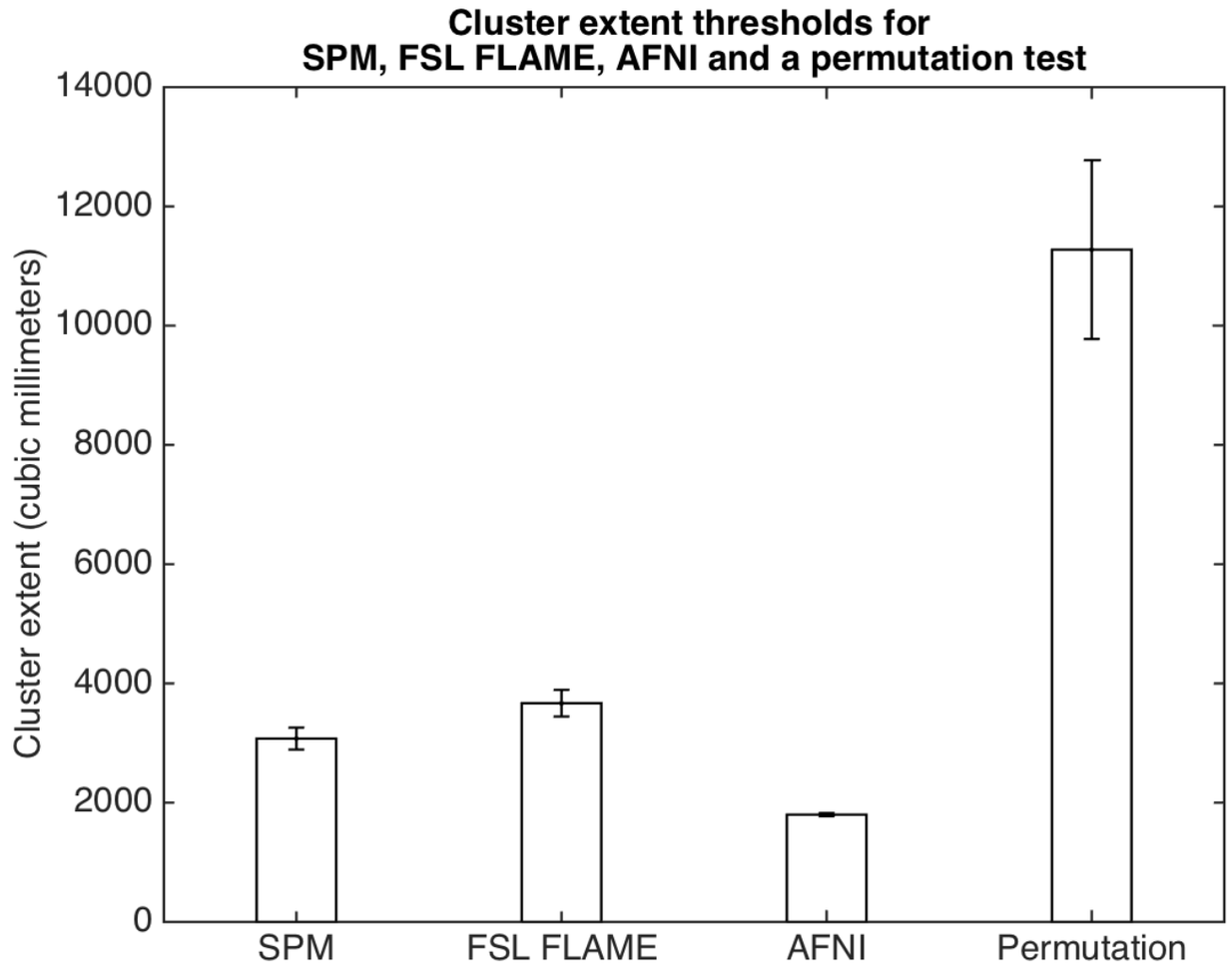
Multiple comparison correction

- Assessed latest 100 papers matching query for fMRI activation studies (circa early 2016)
  - 65 reported whole-brain activation data
  - Good news
    - only 3 papers reported uncorrected results
  - Worrysome news
    - 11% of papers analyzed data using SPM/FSL but then corrected for multiple comparisons using AFNI's alphasim/3dclustsim
  - Why is this a problem?

Poldrack et al, 2017, NRN

P-hacking multiple comparison correction?





Together, the lower group smoothness and the bug in 3dClustSim resulted in cluster extent thresholds that are much lower compared with SPM and FSL (*SI Appendix*, Fig. S16), which resulted in particularly high FWE rates. We find this to be alarming, as 3dClustSim is one of the most popular choices for multiple-comparisons correction (26).

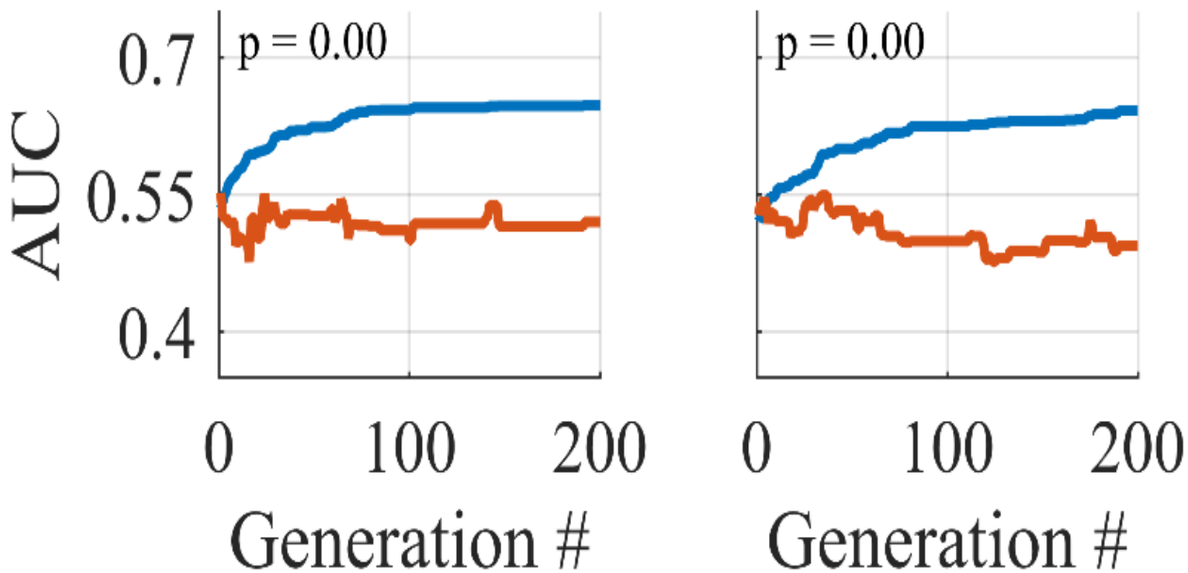
Eklund et al., 2016, *PNAS*

Machine learning can make it worse

I TRIED A BUNCH OF THINGS: THE DANGERS OF UNEXPECTED OVERFITTING IN CLASSIFICATION

MICHAEL SKOCIK[1], JOHN COLLINS[2], CHLOE CALLAHAN-FLINTOFT[3],HOWARD BOWMAN[4], AND BRAD WYBLE[3]

In this article, we use Support Vector Machine (SVM) classifiers, and genetic algorithms to demonstrate the ease by which overfitting can occur, despite the use of cross validation. We demonstrate that comparable and non-generalizable results can be obtained on informative and non-informative (i.e. random) data by iteratively modifying hyperparameters seemingly innocuous ways.



It's not just fMRI

PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 146–157. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12639

How to get statistically significant effects in any ERP experiment (and why you shouldn't)

STEVEN J. LUCK[a,b] AND NICHOLAS GASPELIN[a]

The purpose of this paper is to demonstrate how common and seemingly innocuous methods for quantifying and analyzing ERP effects can lead to very high rates of significant but bogus effects, with the likelihood of obtaining at least one such bogus effect exceeding 50% in many experiments.

Improvement: Pre-registration

- Register sample size and analysis plan up front
  - Preferably with code based on analysis of simulated data
- This does not prevent exploratory analysis
  - But planned and exploratory analyses should be clearly delineated in the paper

http://www.russpoldrack.org/2016/09/why-preregistration-no-longer-makes-me.html

The requirement for clinical trial registration was associated with many more null effects

This is a “cost” under the current incentives to publish

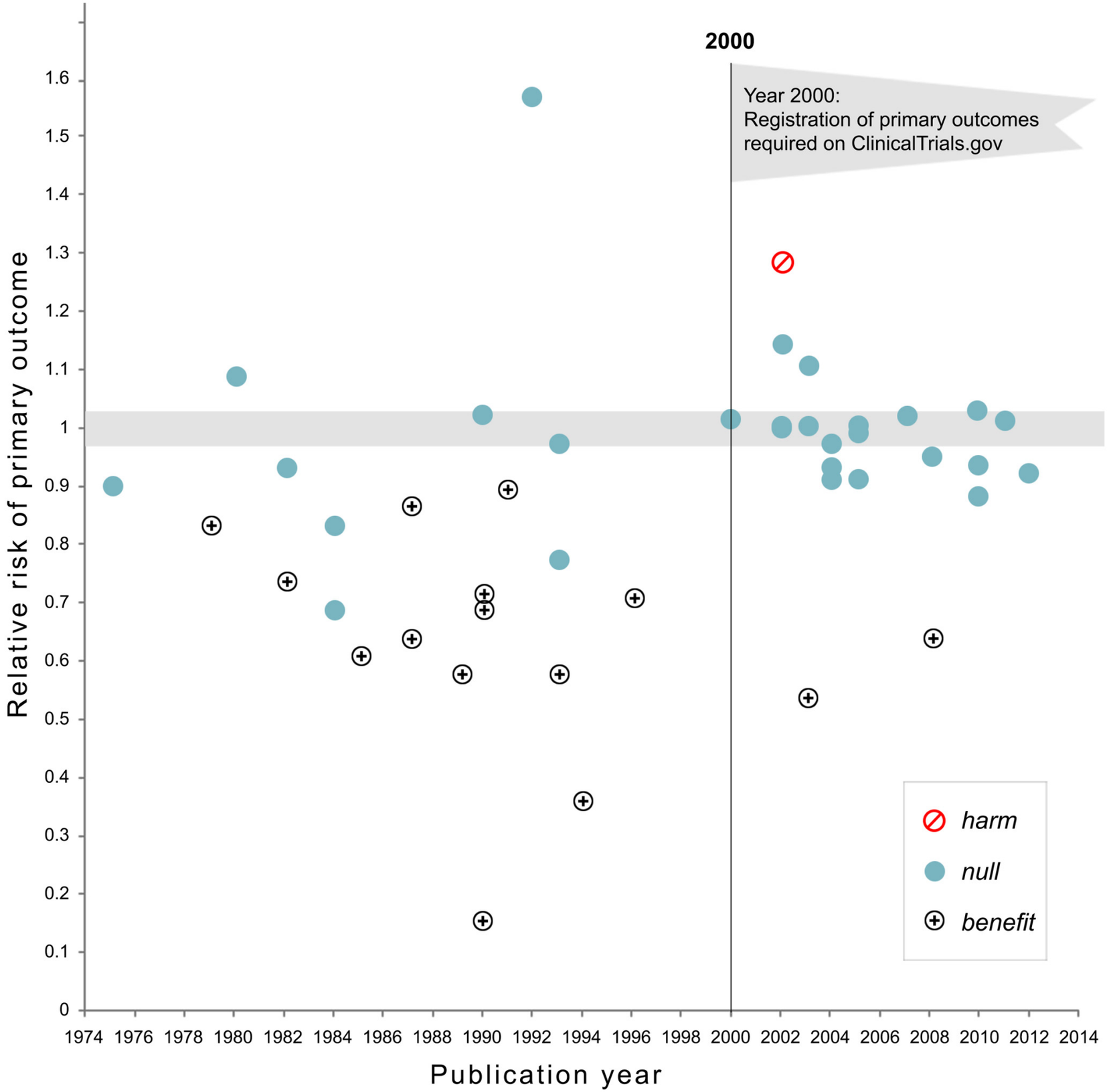


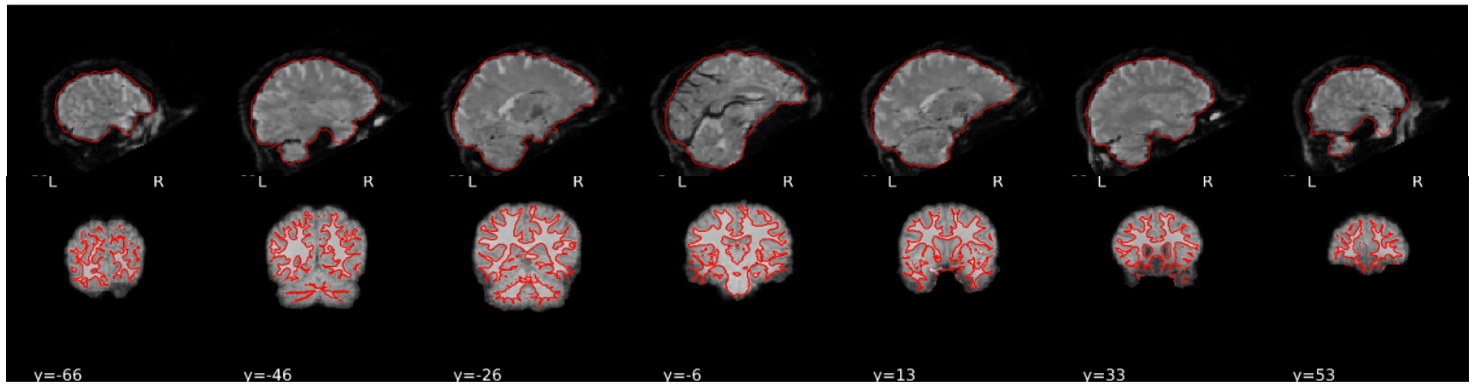
Fig 1. Relative risk of showing benefit or harm of treatment by year of publication for large NHLBI trials on pharmaceutical and dietary supplemen interventions. Positive trials are indicated by the plus signs while trials showing harm are indicated by a diagonal line within a circle. Prior to 2000 when trials were not registered in clinical trials.gov, there was substantial variability in outcome. Following the imposition of the requirement that trials preregister in clinical trials.gov the relative risk on primary outcomes showed considerably less variability around 1.0.

Kaplan & Irvin, 2015

Improvement: Use well-engineered standard tools

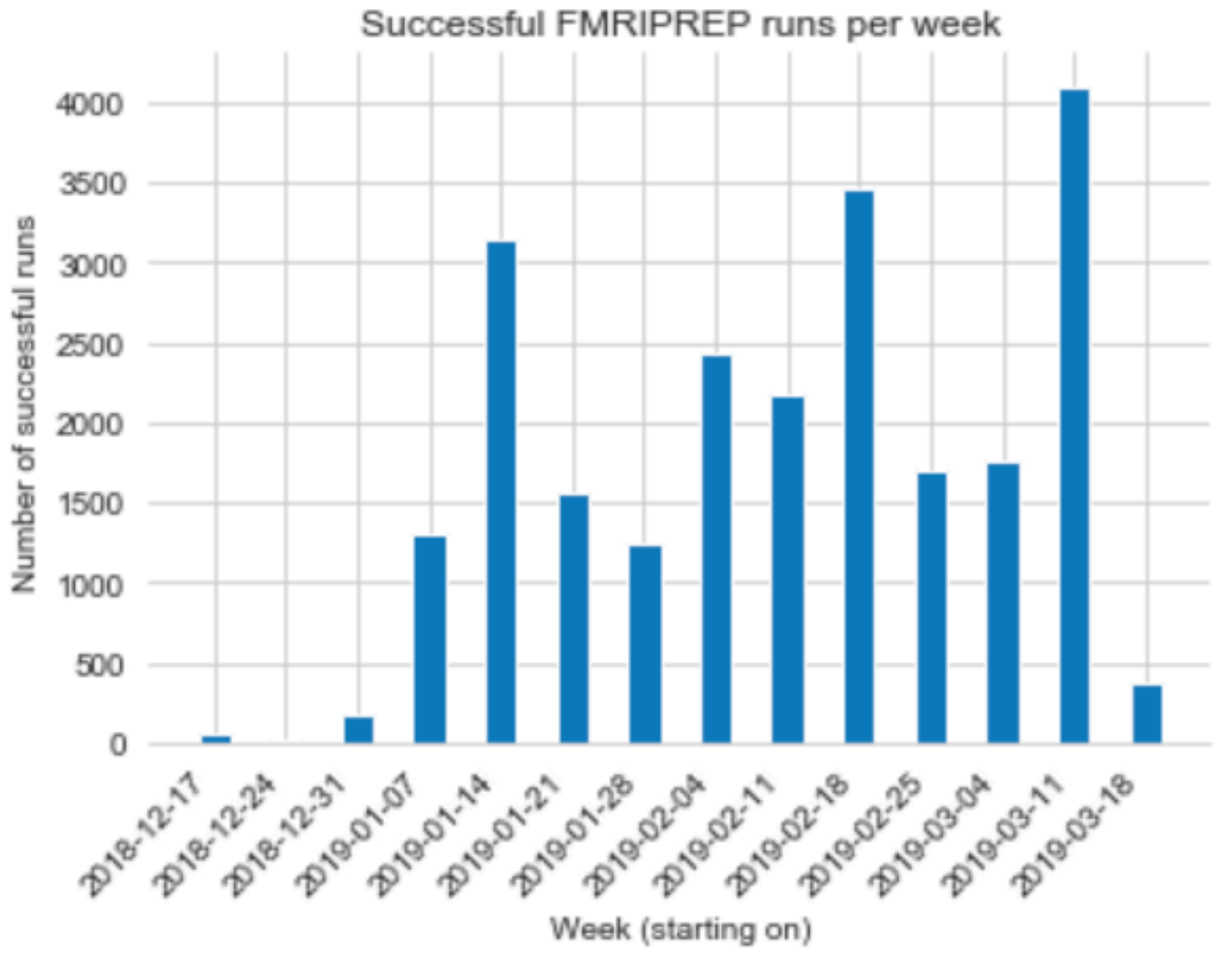
fmriprep: a robust and transparent preprocessing pipeline

Robust
takes any dataset,
combines well tested tools
across packages to provide
the best results



Easy to use
Uses containers, works on Win, Mac,
Linux and HPCs. Takes standardized
datasets (BIDS) and outputs standardized
derivatives.

Transparent
Produces interactive reports
that allow you to check
quality in minutes.



Esteban et al., 2019, *Nature Methoas*

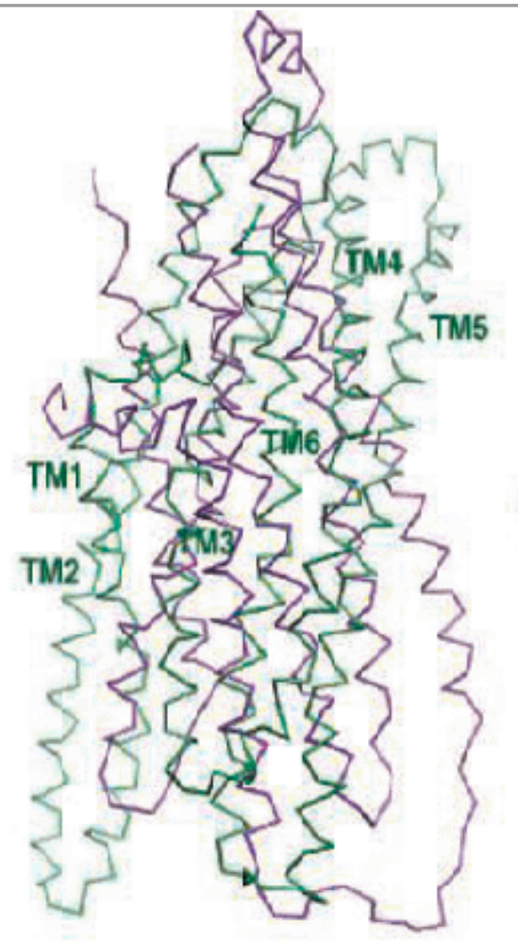
- How many of you have written computer code in the course of your research?

- How many of you have been trained in software engineering?

- How many of you have ever written a test for your code?

Threats to reproducibility: Software errors

Geoffrey Chang



Structure of MsbA from *E. coli*: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters

Geoffrey Chang* and Christopher B. Roth

Multidrug resistance (MDR) is a serious medical problem and presents a major challenge to the treatment of disease and the development of novel therapeutics. ABC transporters that are associated with multidrug resistance (MDR-ABC transporters) translocate hydrophobic drugs and lipids from the inner to the outer leaflet of the cell membrane. To better elucidate the structural basis for the "flip-flop" mechanism of substrate movement across the lipid bilayer, we have determined the structure of the lipid flippase MsbA from *Escherichia coli* by x-ray crystallography to a resolution of 4.5 angstroms. MsbA is organized as a homodimer with each subunit containing six transmembrane α-helices and a nucleotide-binding domain. The asymmetric distribution of charged residues lining a central chamber suggests a general mechanism for the translocation of substrate by MsbA and other MDR-ABC transporters. The structure of MsbA can serve as a model for the MDR-ABC transporters that confer multidrug resistance to cancer cells and infectious microorganisms.

Structure of the ABC Transporter MsbA in Complex with ADP·Vanadate and Lipopolysaccharide

Christopher L. Reyes and Geoffrey Chang*

Select members of the adenosine triphosphate (ATP)–binding cassette (ABC) transporter family couple ATP binding and hydrolysis to substrate efflux and confer multidrug resistance. We have determined the x-ray structure of MsbA in complex with magnesium, adenosine diphosphate, and inorganic vanadate ($Mg{\cdot}ADP{\cdot}V_i$) and the rough-chemotype lipopolysaccharide, Ra LPS. The structure supports a model involving a rigid-body torque of the two transmembrane domains during ATP hydrolysis and suggests a mechanism by which the nucleotide-binding domain communicates with the transmembrane domain. We propose a lipid "flip-flop" mechanism in which the sugar groups are sequestered in the chamber while the hydrophobic tails are dragged through the lipid bilayer.

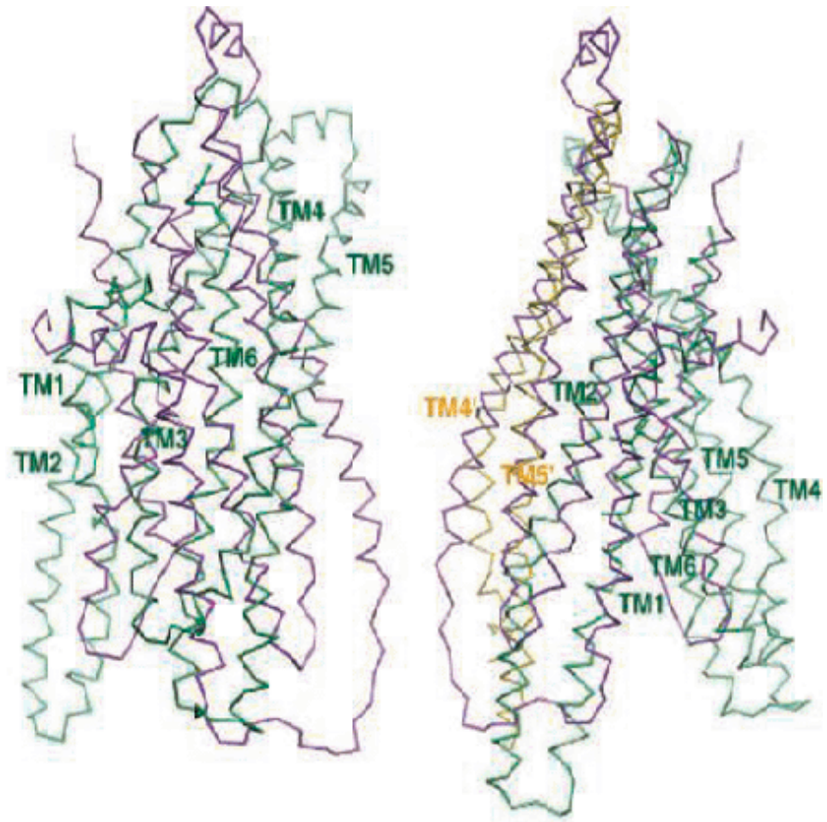
X-ray Structure of the EmrE Multidrug Transporter in Complex with a Substrate

Owen Pornillos, Yen-Ju Chen, Andy P. Chen, Geoffrey Chang*

EmrE is a prototype of the Small Multidrug Resistance family of efflux transporters and actively expels positively charged hydrophobic drugs across the inner membrane of *Escherichia coli*. Here, we report the x-ray crystal structure, at 3.7 angstrom resolution, of one conformational state of the EmrE transporter in complex with a translocation substrate, tetraphenylphosphonium. Two EmrE polypeptides form a homodimeric transporter that binds substrate at the dimerization interface. The two subunits have opposite orientations in the membrane and adopt slightly different folds, forming an asymmetric antiparallel dimer. This unusual architecture likely confers unidirectionality to transport by creating an asymmetric substrate translocation pathway. On the basis of available structural data, we propose a model for the proton-dependent drug efflux mechanism of EmrE.

Threats to reproducibility: Software errors

Retraction



WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP•vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

Small errors can have big effects

```
# 23-class classification problem

skf=StratifiedKFold(labels,8)

if trainsvm:
    pred=N.zeros(len(labels))
    for train,test in skf:
        clf=LinearSVC()
        clf.fit(data[train],labels[train])
        pred[test]=clf.predict(data[test])
```

data[:,train]

data[:,test]

Results:
93% accuracy

Results:
53% accuracy

http://www.russpoldrack.org/2013/02/anatomy-of-coding-error.html

Bug-hacking

- Bugs that confirm our predictions are less likely to be uncovered than bugs that disconfirm them

The principle of assumed error

- Whenever you find a seemingly good result (e.g. one that fits your predictions), assume that it occurred due to an error in your code
- Protects from “bug hacking”

http://www.russpoldrack.org/2016/08/the-principle-of-assumed-error.html

Improvement: Software testing

- Smoke tests
  - Does everything run without exploding?
- Sanity checks
  - Are values within reasonable ranges?

PLOS ONE

CORRECTION

Correction: The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science

Stephan Lewandowsky, Gilles E. Gignac, Klaus Oberauer

The dataset included two notable age outliers (reported ages 5 and 32757).

Specifically, the statement on page 9 “age turned out not to correlate with any of the indicator variables” is incorrect. It should read instead “age correlated significantly with 3 latent indicator variables (Vaccinations: .219, $p < .0001$; Conservatism: .169, $p < .001$; Conspiracist ideation: -.140, maximum likelihood $p < .0001$, bootstrapped $p = .004$), and straddled significance for a fourth (Free Market: .08, p%.05).”

```
In [1]: age=32757

In [2]: assert age>12 and age<120
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-2-37de876b5fda> in <module>()
----> 1 assert age>12 and age<120

AssertionError:
```

Automating quality control using continuous integration

- Free automated testing platforms for open source projects (CircleCI, Travis)
- Automatically runs software tests whenever a new commit is pushed to Github

Improvement: Code quality analysis

- Static analysis (pylint, flake8, etc)
  - Checks for errors and adherence to Python style guidelines
  - Collaborator: “Fixing the style made your code so much easier to read!”

```
$ pylint cards.py
************* Module python_cards.cards
W:  4, 0: Found indentation with tabs instead of spaces (mixed-indentation)
C:  1, 0: Missing module docstring (missing-docstring)
C:  3, 0: Missing class docstring (missing-docstring)
C:  3, 0: Old-style class defined. (old-style-class)
R:  3, 0: Too few public methods (0/2) (too-few-public-methods)
W:  9,20: Redefining built-in 'list' (redefined-builtin)
```

https://www.codacy.com/blog/which-python-static-analysis-tools-should-i-use/

Improvement: Use established libraries when possible

- Avoid the NIH ("not invented here") effect
  - rejecting existing solutions in favor of home-grown ones
  - "I need to write a new DICOM to Nifti converter"
- Contribute fixes/extensions to existing open source projects rather than writing your own
- Prefer libraries that use good software engineering practices

scikit-learn

scikit-learn is a Python module for machine learning built on top of SciPy and distributed under the 3-Clause BSD license.

Creating reproducible analyses: A case study

- Neuroimaging Analysis Replication and Prediction Study (NARPS)
  - 70 teams analyzed a shared fMRI dataset and submitted hypothesis tests and maps
  - Results showed substantial variability in hypothesis testing outcomes across teams
  - Analysis of submitted maps required a complex custom workflow

https://github.com/poldrack/narps

http://narps.info

Botvinik-Nezer et al, in prep

Steps towards reproducibility: Data versioning

"The full dataset is publicly available on OpenNeuro (DOI: 10.18112/openneuro.ds001734.v1.0.4). "

SCIENTIFIC DATA

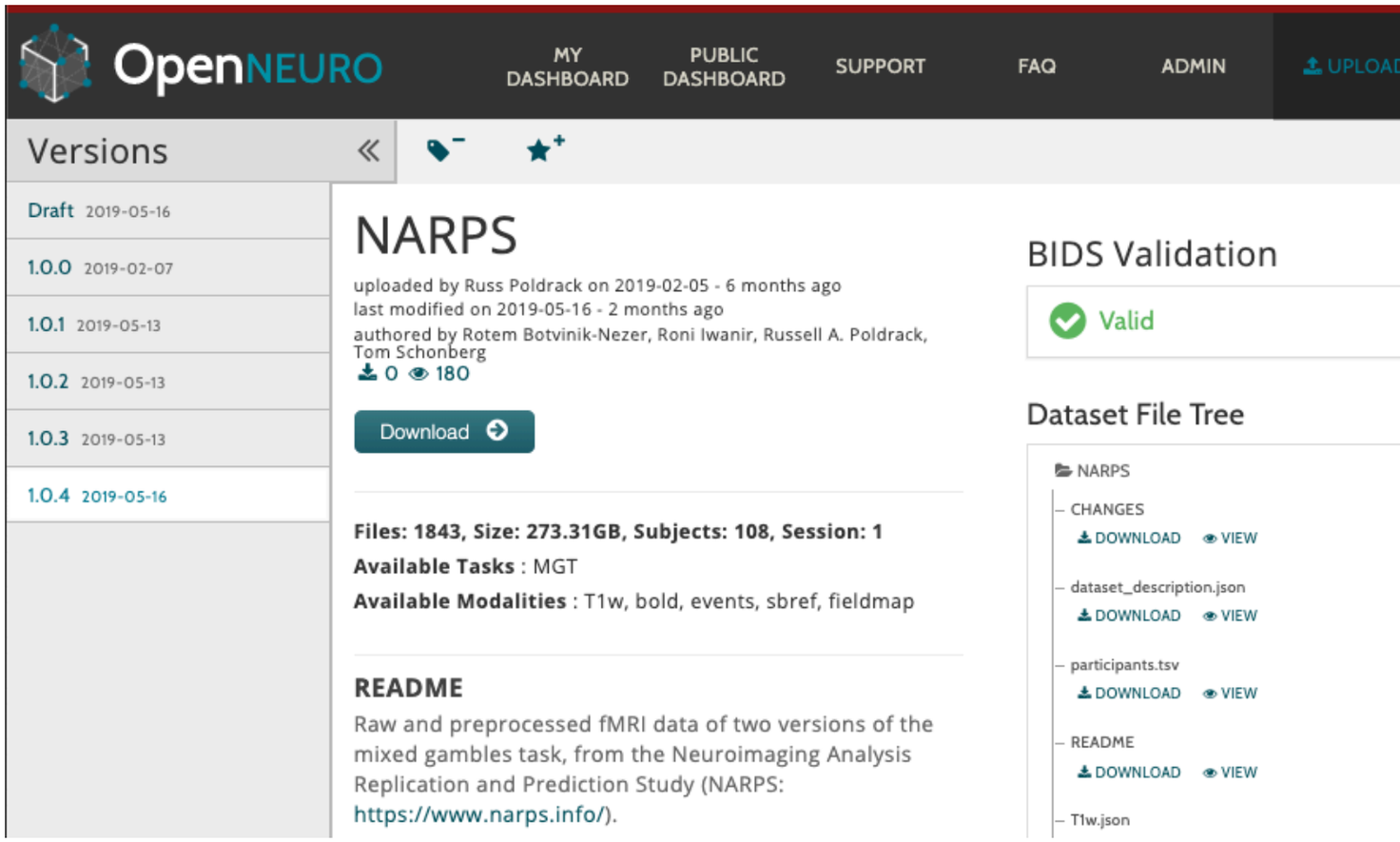
OPEN

DATA DESCRIPTOR

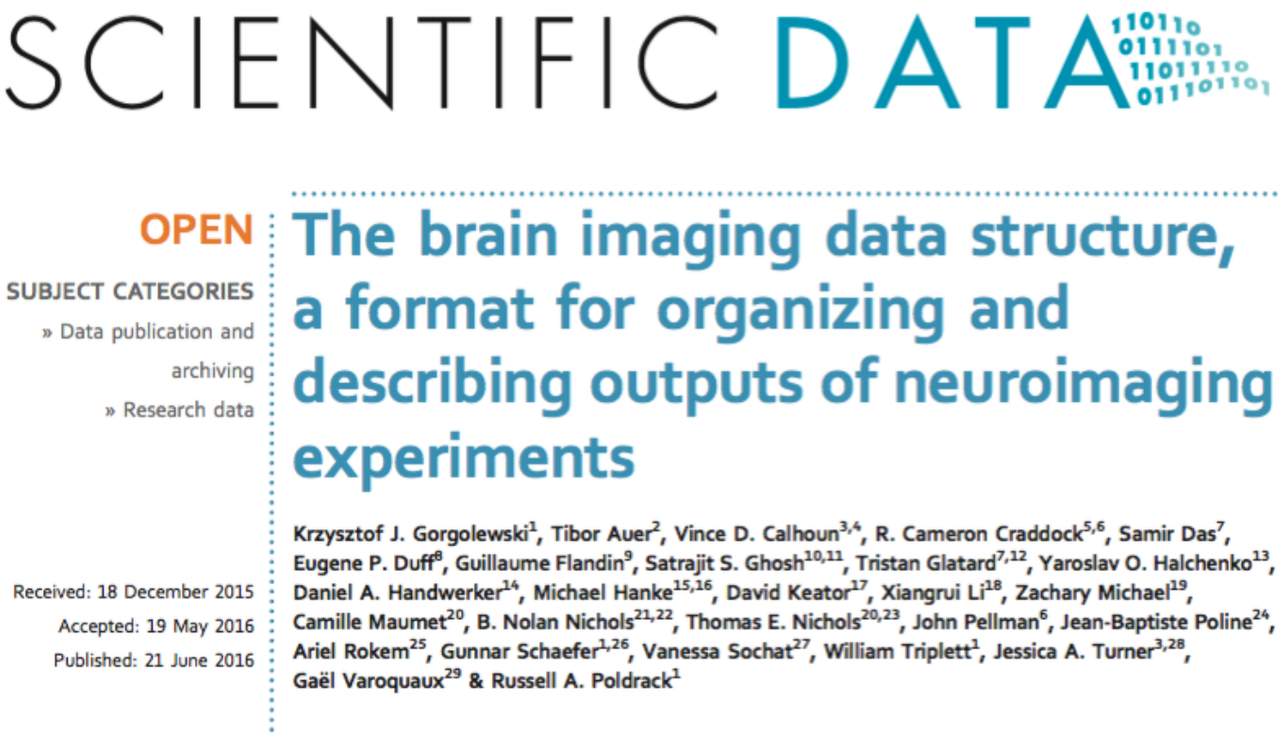
fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study

Received: 18 February 2019
Accepted: 7 June 2019
Published online: 01 July 2019

Rotem Botvinik-Nezer[1,2], Roni Iwanir[1,2], Felix Holzmeister[3], Jürgen Huber[3], Magnus Johannesson[4], Michael Kirchler[3], Anna Dreber[4,5], Colin F. Camerer[6], Russell A. Poldrack[7] & Tom Schonberg[1,2]



The importance of standards for data sharing


SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Data publication and archiving
» Research data

The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

Krzysztof J. Gorgolewski[1], Tibor Auer[2], Vince D. Calhoun[3,4], R. Cameron Craddock[5,6], Samir Das[7], Eugene P. Duff[8], Guillaume Flandin[9], Satrajit S. Ghosh[10,11], Tristan Glatard[7,12], Yaroslav O. Halchenko[13], Daniel A. Handwerker[14], Michael Hanke[15,16], David Keator[17], Xiangrui Li[18], Zachary Michael[19], Camille Maumet[20], B. Nolan Nichols[21,22], Thomas E. Nichols[20,23], John Pellman[6], Jean-Baptiste Poline[24], Ariel Rokem[25], Gunnar Schaefer[1,26], Vanessa Sochat[27], William Triplett[1], Jessica A. Turner[3,28], Gaël Varoquaux[29] & Russell A. Poldrack[1]

Received: 18 December 2015
Accepted: 19 May 2016
Published: 21 June 2016

- sub-control01
  - anat
    - sub-control01_T1w.nii.gz
    - sub-control01_T1w.json
    - sub-control01_T2w.nii.gz
    - sub-control01_T2w.json
  - func
    - sub-control01_task-nback_bold.nii.gz
    - sub-control01_task-nback_bold.json
    - sub-control01_task-nback_events.tsv
    - sub-control01_task-nback_physio.tsv.gz
    - sub-control01_task-nback_physio.json
    - sub-control01_task-nback_sbref.nii.gz
  - dwi
    - sub-control01_dwi.nii.gz

- Current standards exist for:
  - s/f/dMRI, MEG, EEG/iEEG
- In the works for:
  - Modalities: PET, ASL, eye tracking, MRS
  - Derivatives, statistical/computational models, spatial transforms, genetics

http://bids.neuroimaging.io

Steps toward reproducibility: Containerization

- "an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." - Buckheit & Donoho, 1995
- Containers provide a way to share an entire computing platform
  - Including specific versions of all software and libraries


```
FROM python:3.6-stretch

# apt-get installs

RUN apt-get update && apt-get install -y --no-install-recommends \
    gcc=4:6.3.0-4 \
    vim=2:8.0.0197-4+deb9u3 \
    wget=1.18-5+deb9u3

# pip installs

RUN pip install \
    nibabel==2.4.1 \
    nipype==1.2.0 \
    nilearn==0.5.2 \
```

Steps towards reproducibility: R

- Pinning versions of R code is tricky
- We used the **`checkpoint`** library
  - Downloads all libraries as of a specific date
  - Greatly slows execution if many libraries are used

```
> checkpoint(“2019-07-16")

checkpoint: Part of the Reproducible R Toolkit from
Microsoft
https://mran.microsoft.com/documents/rro/reproducibility/
Scanning for packages used in this project
- Discovered 14 packages
Installing packages used in this project
 - Installing ‘arm’…
```

Steps toward reproducibility: Makefile

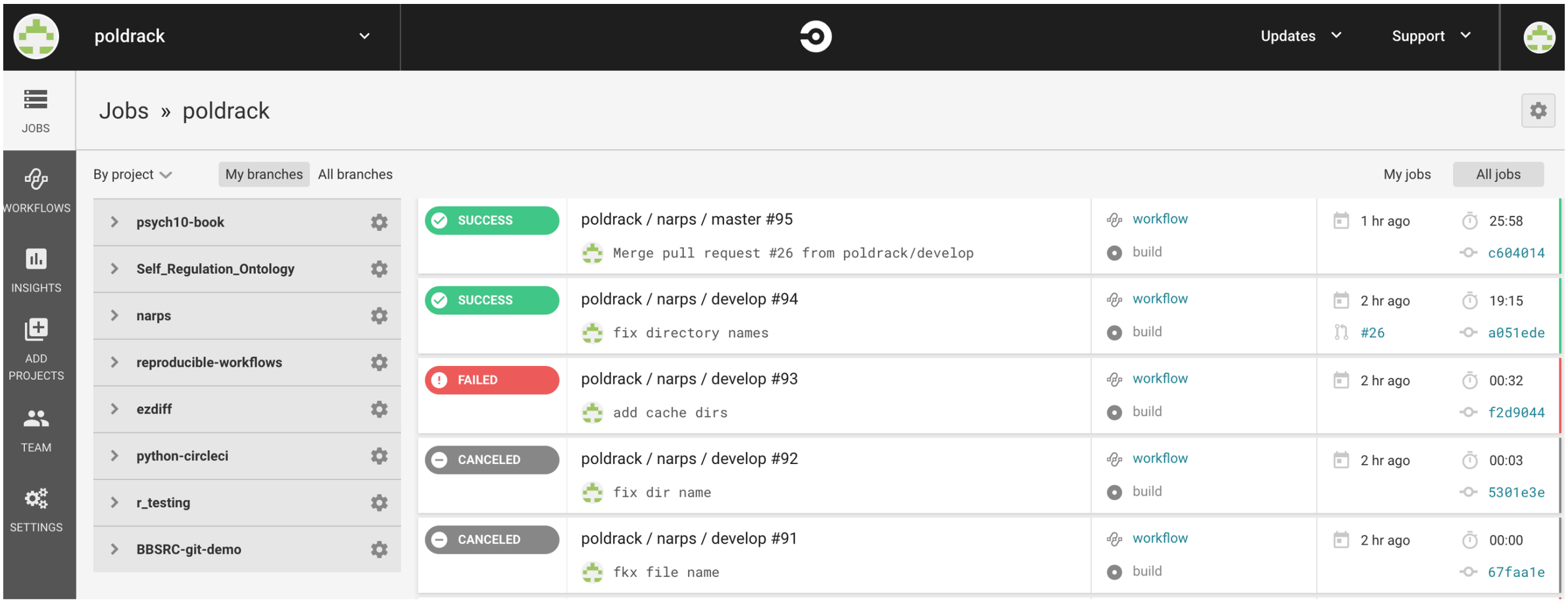
- Useful for users
  - `make run-all`
- Instead of:
  - `docker run -e "DATA_URL=$(DATA_URL)" -v $(current_dir):/analysis -v $(NARPS_BASEDIR):/data $(DOCKER_USERNAME)/narps-analysis /bin/bash -c "source /etc/fsl/5.0/fsl.sh;python /analysis/narps.py"`
- Useful for you as a developer
  - `make check-style`

Steps toward reproducibility: Automated testing

- CircleCI config file (.circleci/config.yml)

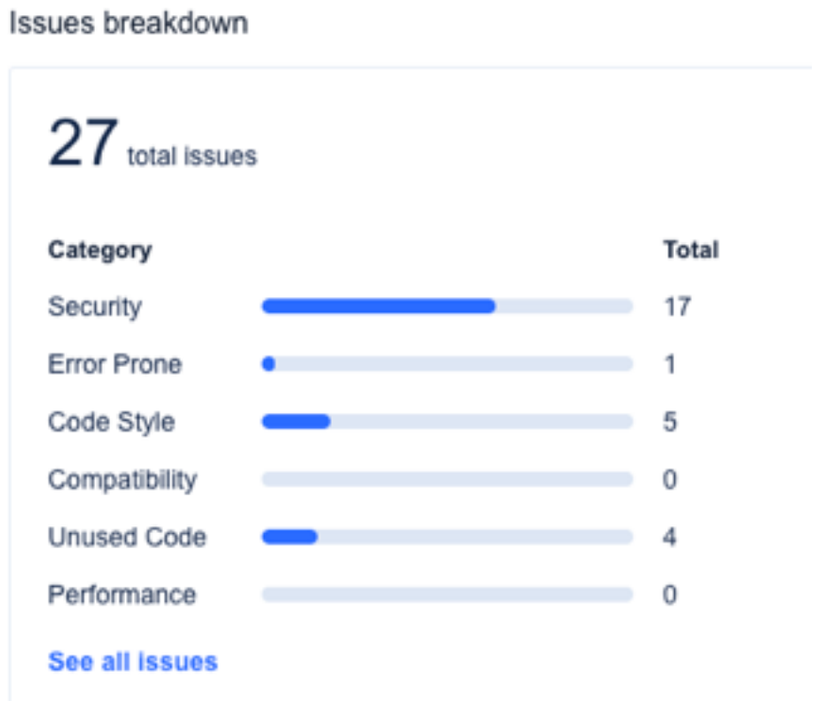
```
version: 2
jobs:
  build:
    docker:
      - image: poldrack/narps-analysis
    steps:
      - checkout
      - run:
          name: run main narps tests
          command: |
            pip install -U pytest pytest-cov
            source /etc/fsl/5.0/fsl.sh; pytest --cov=./ImageAnalyses-q
ImageAnalyses/tests.py
      - run:
          name: static analysis and style check using flake8
          command: |
            pip install -U flake8
            flake8 ImageAnalyses
      - run:
          name: coverage
          command: |
             pip install coveralls
             coveralls
```

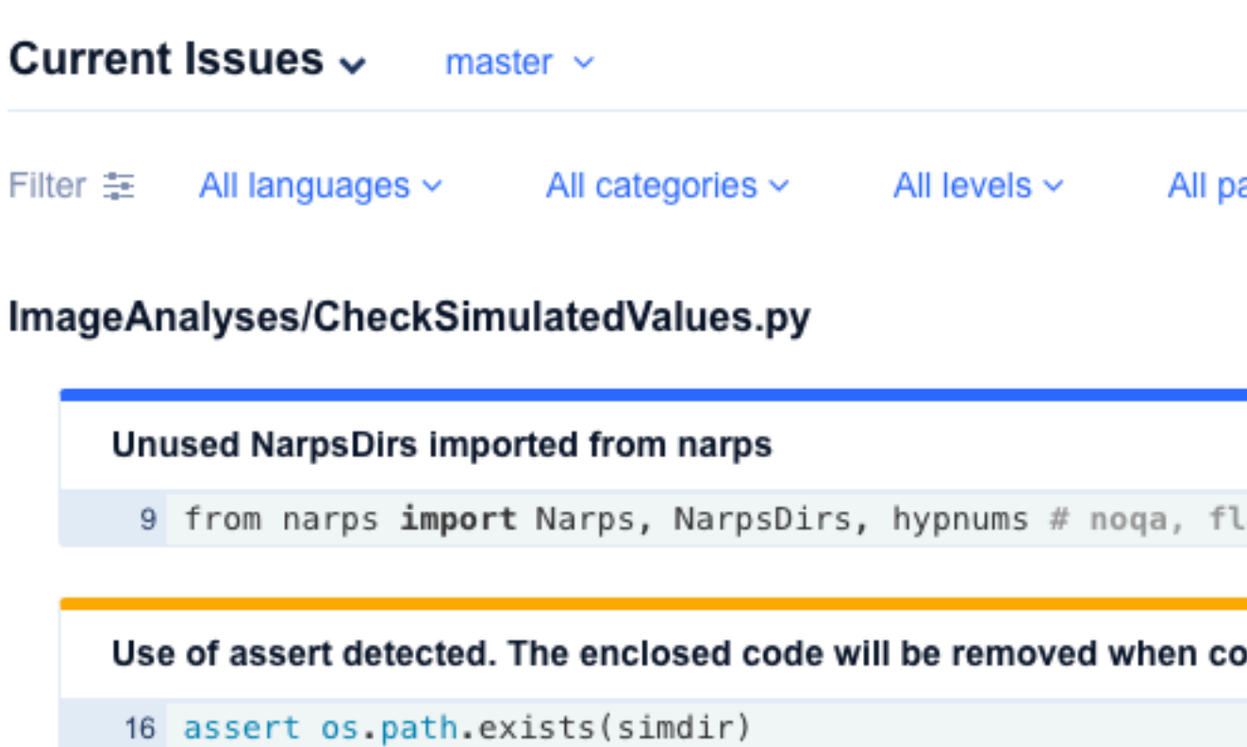
Automated testing using CircleCI



Automated code quality assessment (codacy.com)

- Automatically checks code upon Github push





Test coverage using coveralls

POLDRACK / NARPS / 88

COMMITTED 23 JUL 2019 - 14:13 — COVERAGE REMAINED THE SAME AT 68.527%

SOURCE FILES ON BUILD 88

TREE | LIST 11 | CHANGED 0 | SOURCE CHANGED 0 | COVERAGE CHANGED 0

68.53	analysis/
74.91	AnalyzeMaps.py
0.0	CheckSimulatedValues.py
70.37	ConsensusAnalysis.py
74.29	MakeSupplementaryFigure1.py
0.0	PrepareData.py
0.0	PrepareMapDiagnosticReports.py
65.85	PrepareMetadata.py
79.65	narps.py
100.0	tests.py

Steps toward reproducibility: Code review

- Enlisted an individual with software engineering and imaging analysis expertise to review the code
  - Primary goal: Does the logic make sense?
  - Secondary goal: Did I make any stupid little mistakes?

Snapshotting and DOI generation

- Integration between Github and Zenodo allows generation of a Digital Object Identifier for each release of a project
- This allows specification of exactly what code went into an analysis

Badges!

Neuroimaging Analysis Replication and Prediction Study

This repository contains code related to the Neuroimaging Analysis Replication and Prediction Study.

For analyses of the prediction market data, see the PredictionMarketAnalyses directory.

For analyses of the imaging results, see the ImageAnalyses directory.

Steps toward reproducibility: Analytic validation

- A result can be fully reproducible but incorrect
  - cf. distinction between reliability and validity
- For complex analyses:
  - Parameter recovery: Generate data for which the true answer is known, and assess ability of code to recover the correct answer
  - Randomization: Generate data for which the null hypothesis of no relationship should be true on average, and ensure that the observed false positive rate is accurate (cf. Eklund et al., 2016, PNAS)
- Best done in advance of any real analyses and pre-registered

http://www.russpoldrack.org/2016/08/the-principle-of-assumed-error.html

Validation for NARPS study

- Generate simulated data for each analysis team
  - Generate simulated data based on known maps
  - Include some teams with either completely random data or anticorrelated signal
  - Include different levels of variance across teams
- Run entire analysis stream on the simulated data
- Compare results of final analysis to ground-truth maps used to generate the data

```
SUCESSS: all hypotheses correlated > 0.95
```

Conclusion

- Neuroimaging has a reproducibility problem
  - Likely not very different from other areas of science with complex computational workflows and large datasets
- Statistical power is key to solving the problem
- Software engineering tools can help ensure reproducibility of complex analyses
- The tools that you will learn this week will help you do better science!

Acknowledgments

The Poldrack Lab @ Stanford
http://reproducibility.stanford.edu

Many awesome collaborators
(too many to name!)

NIMH National Institute of Mental Health

NIH National Institute on Drug Abuse

ljaf laura and john arnold foundation®

NSF