

Reproducibility in fMRI: What is the problem?

Russell Poldrack
Stanford University

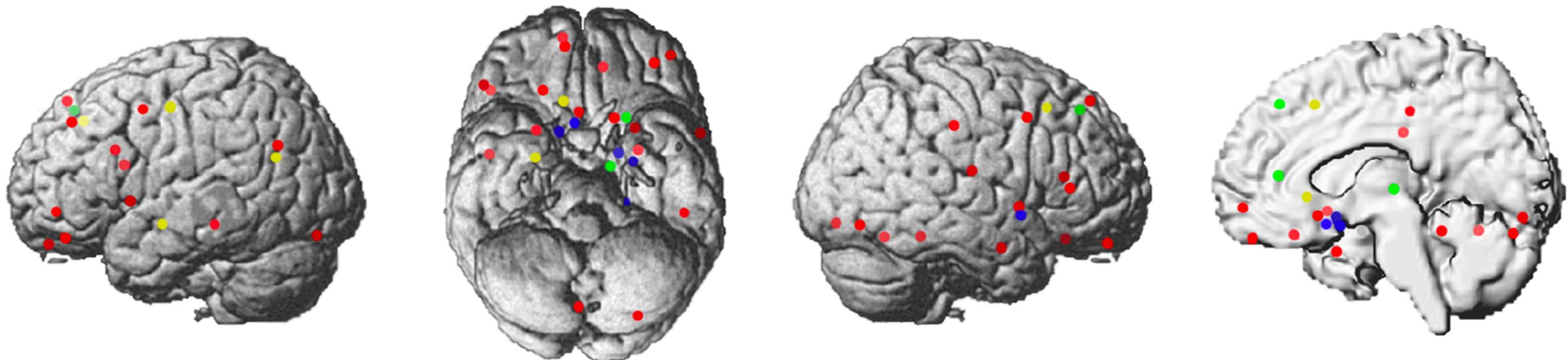
A search for “fMRI” for FY2020 NIH grants finds:

1218 matching grants

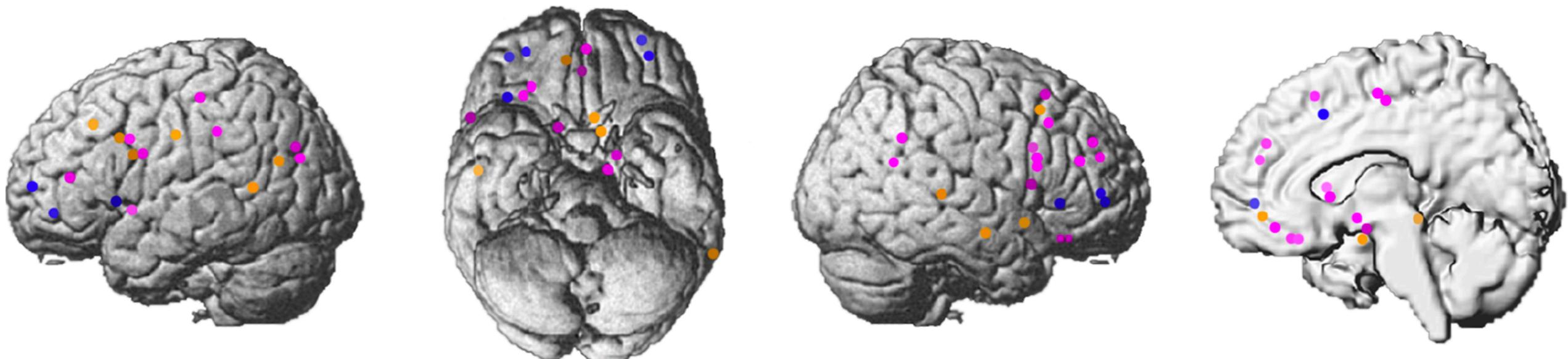
\$552M total costs

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

A Peak coordinates of emotional processing meta-analyses only



B Peak coordinates of cognitive processing and mixed meta-analyses



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, MSc¹; 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-

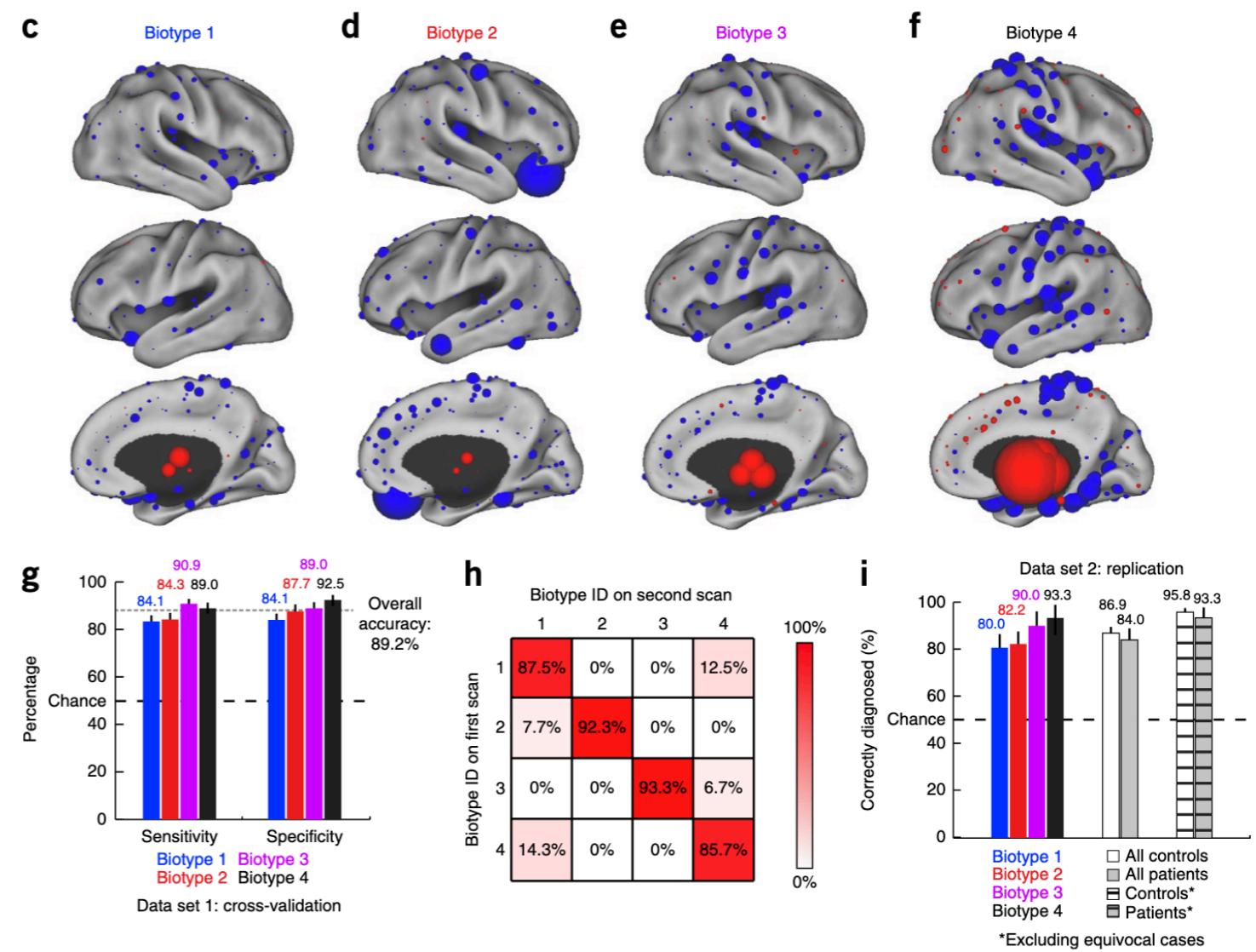
Resting-state connectivity biomarkers define neurophysiological subtypes of depression

Andrew T Drysdale¹⁻³, Logan Grosenick^{4,5}, Jonathan Downar⁶, Katharine Dunlop⁶, Farrokh Mansouri⁶, Yue Meng¹, Robert N Fethcho¹, Benjamin Zebley⁷, Desmond J Oathes⁸, Amit Etkin^{9,10}, Alan F Schatzberg⁹, Keith Sudheimer⁹, Jennifer Keller⁹, Helen S Mayberg¹¹, Faith M Gunning^{2,12}, George S Alexopoulos^{2,12}, Michael D Fox¹³, Alvaro Pascual-Leone¹³, Henning U Voss¹⁴, BJ Casey¹⁵, Marc J Dubin^{1,2} & Conor Liston¹⁻³

VOLUME 23 | NUMBER 1 | JANUARY 2017

NATURE MEDICINE

“Our results define novel subtypes of depression that transcend current diagnostic boundaries and may be useful for identifying the individuals who are most likely to benefit from targeted neurostimulation therapies.”



Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017)

Richard Dinga^{a,*}, Lianne Schmaal^{b,c}, Brenda W.J.H. Penninx^a, Marie Jose van Tol^d,
Dick J. Veltman^a, Laura van Velzen^a, Maarten Mennes^e, Nic J.A. van der Wee^f,
Andre F. Marquand^e

[NeuroImage: Clinical 22 \(2019\) 101796](#)

Here, we aimed to replicate the analytical pipeline proposed by Drysdale and colleagues (Drysdale et al., 2017) and extend their findings to a different clinical population. We followed the analysis steps of the original study as closely as possible and obtained extremely high canonical correlations (0.99 and 0.97) and optimal 3 cluster solution. However, after performing additional tests that were not performed in the original study, which took into account that the RS-fMRI connectivity features were selected based on their correlation with clinical features before performing CCA on these connectivity and clinical features, even the high canonical correlations that we observed were not statistically significant and they did not replicate outside of the training set. By using the same criteria for selecting the number of clusters as in the original study, we found an optimal three cluster solution. However, we showed that this cluster solution would happen even if the data came only from a single Gaussian distribution with no underlying clusters.



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

Stanford | Center for Reproducible Neuroscience

Open data



OpenNEURO^{BETA}

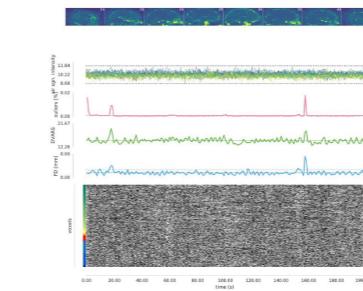
A free and open platform for analyzing
and sharing neuroimaging data



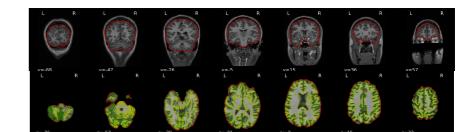
NEUROVAULT

A public repository of unthresholded statistical maps,
parcellations, and atlases of the human brain

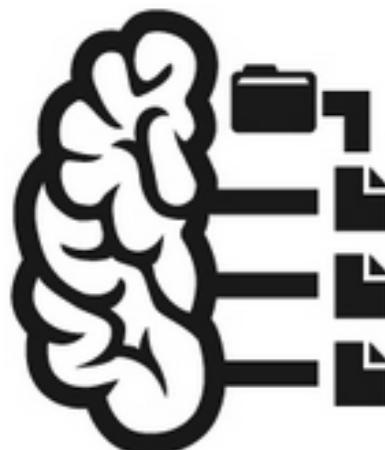
MRIQC



fMRIPrep



Open standards

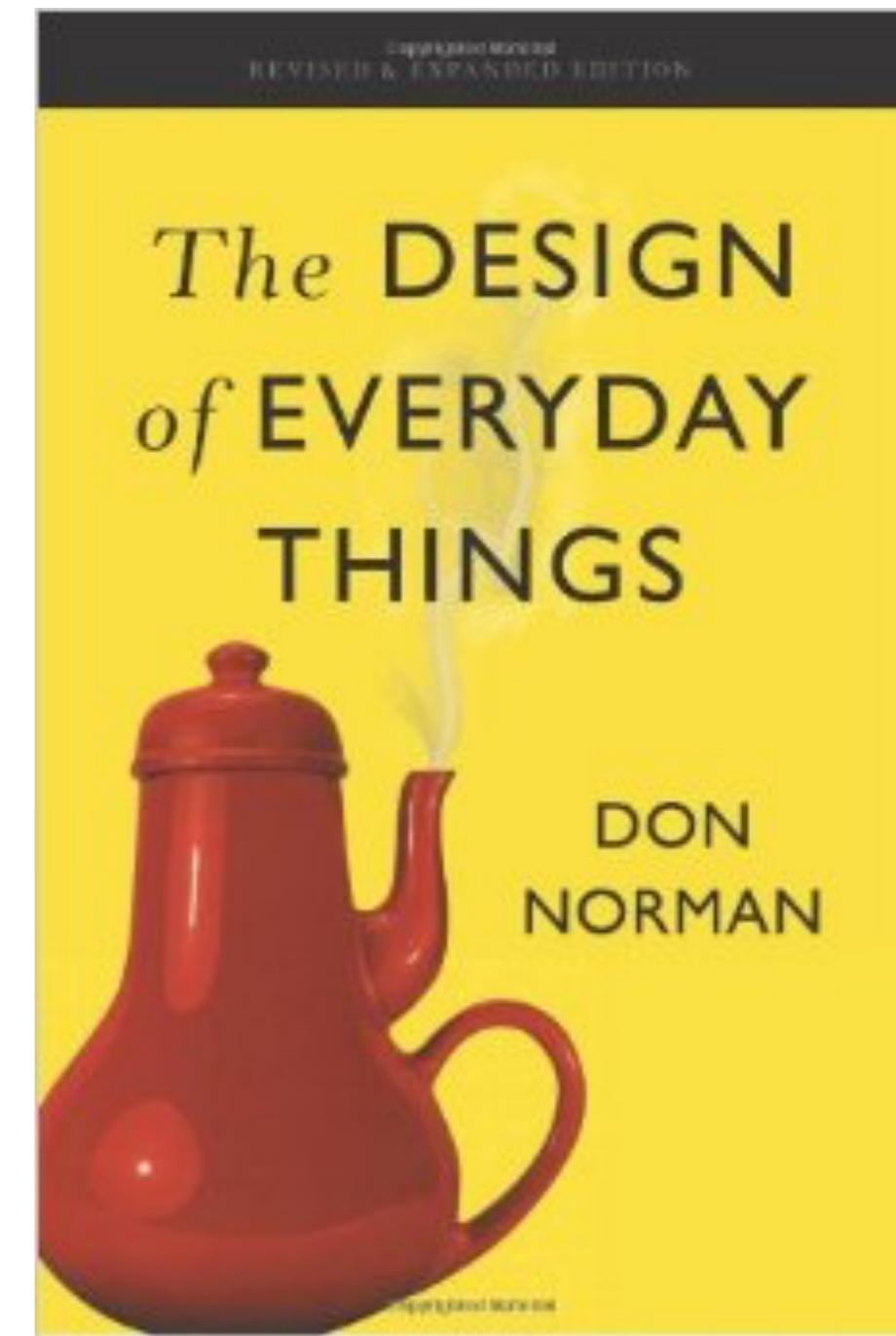


BIDS
BRAIN IMAGING DATA STRUCTURE

Meta-science

#narps: Neuroimaging Analysis
Replication and Prediction Study

Designing a more reproducible scientific enterprise



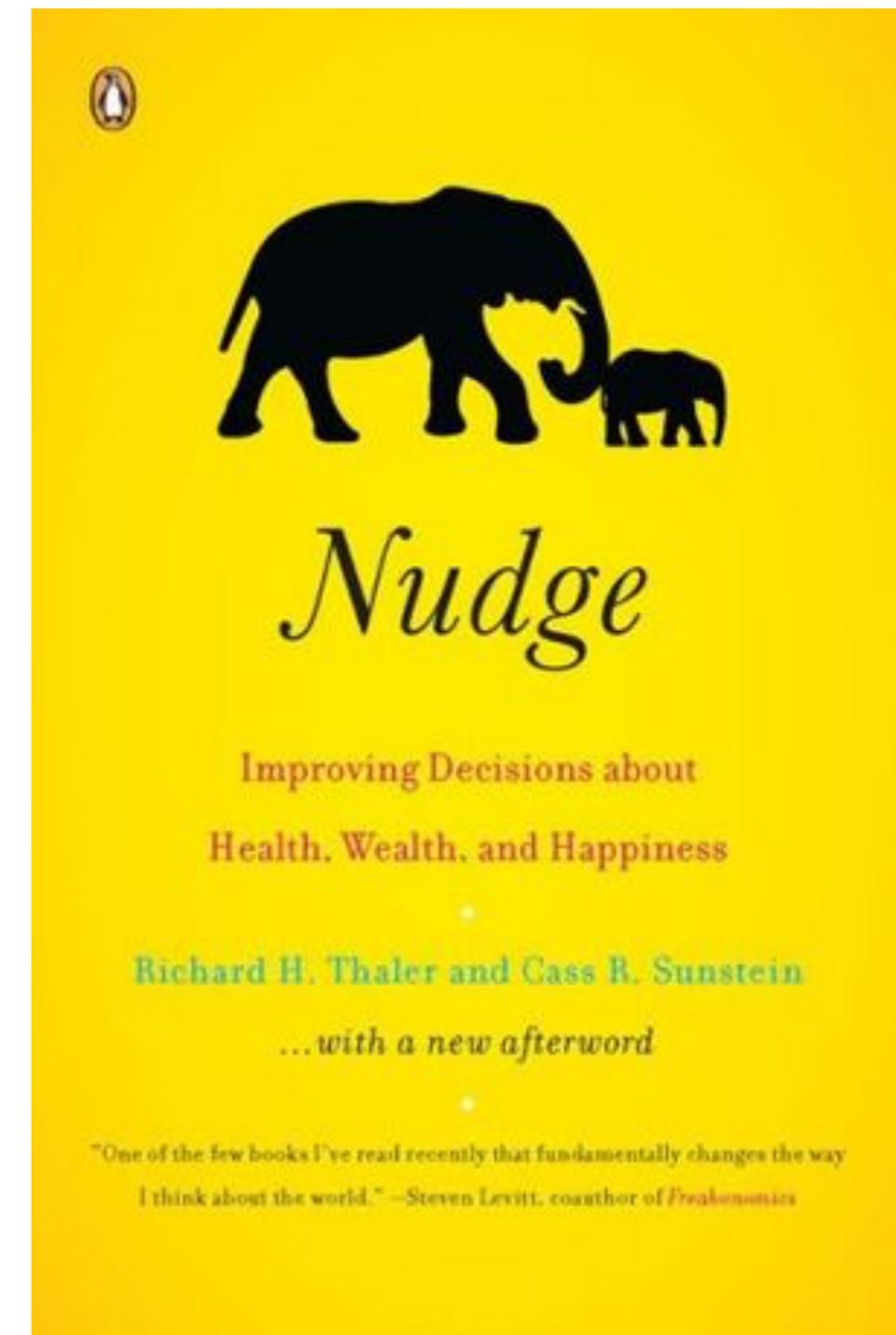


FOOLING OURSELVES

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



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis



PLoS Medicine | www.plosmedicine.org

0696

August 2005 | Volume 2 | Issue 8 | e124

- 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.

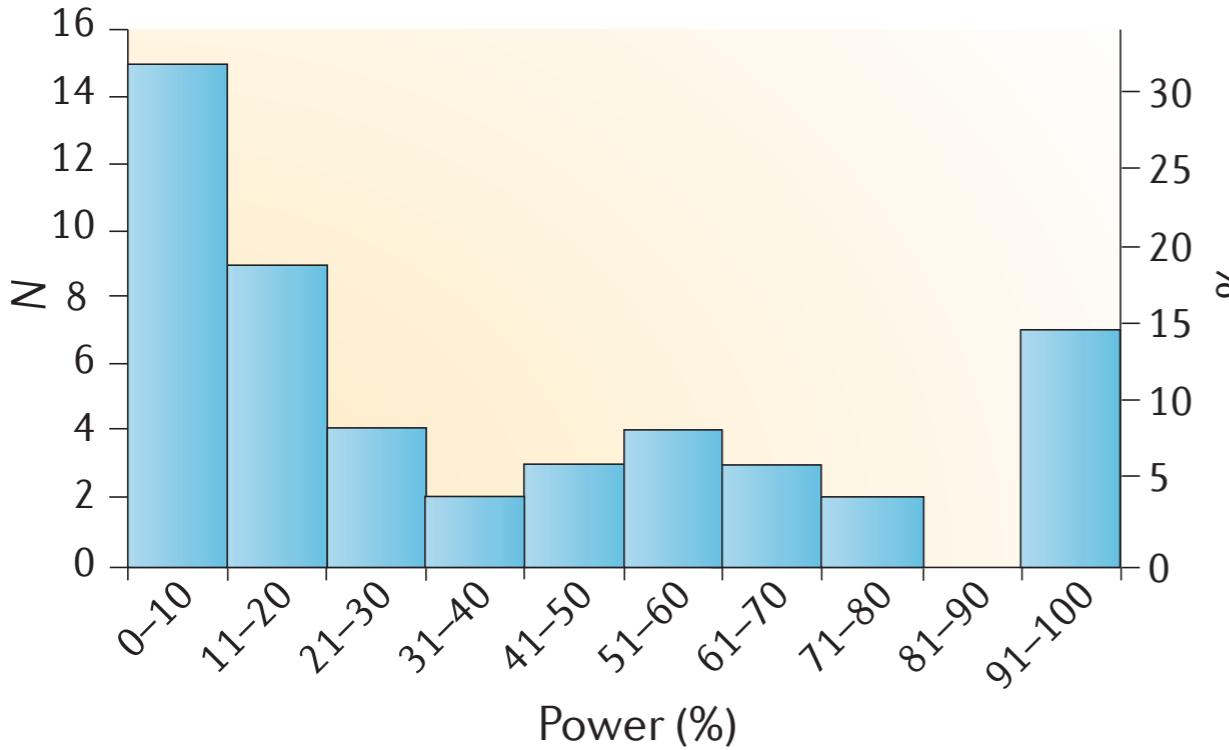
Neuroscience research is badly underpowered

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

Katherine S. Button^{1,2}, John P. A. Ioannidis³, Claire Mokrysz¹, Brian A. Nosek⁴, Jonathan Flint⁵, Emma S. J. Robinson⁶ and Marcus R. Munafò¹

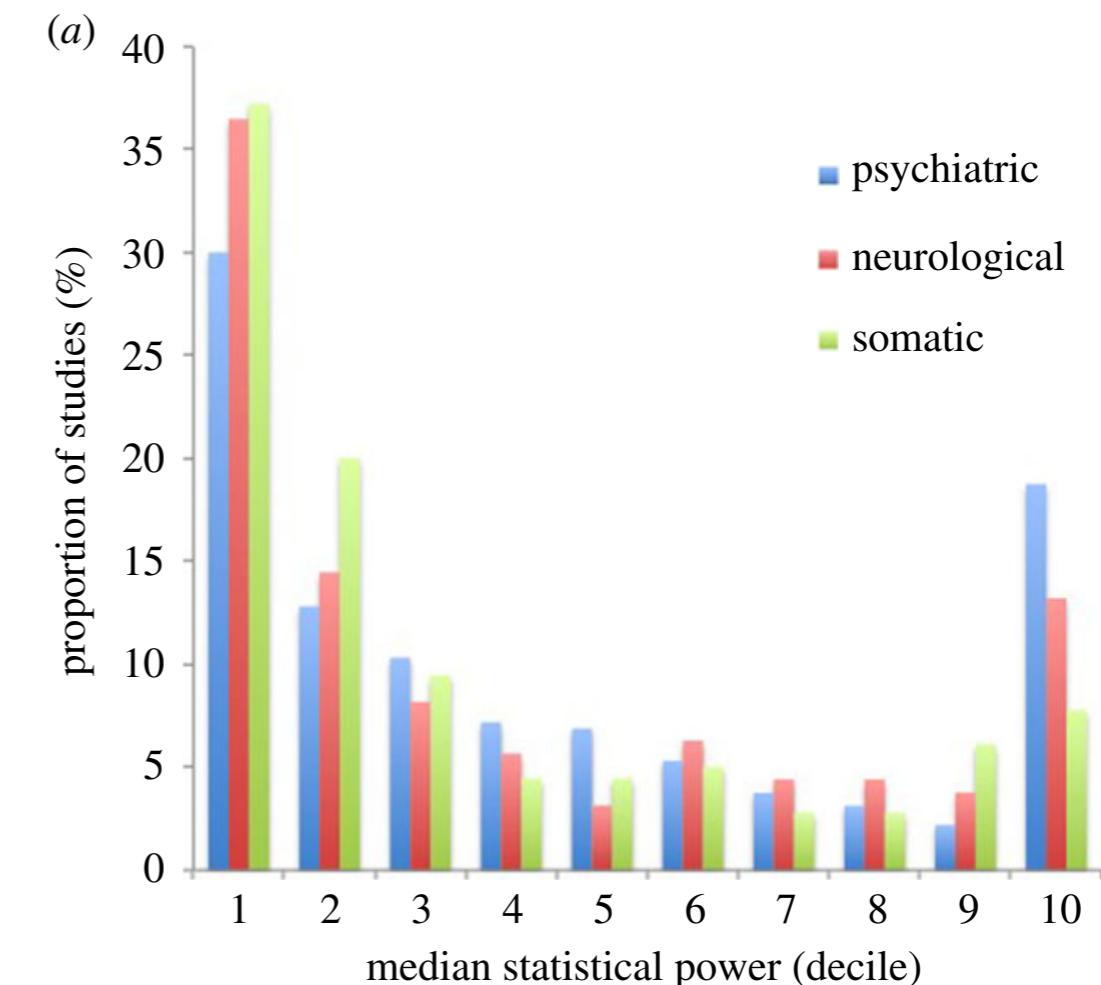
NATURE REVIEWS | NEUROSCIENCE

VOLUME 14 | MAY 2013 | 365



Low statistical power in biomedical science: a review of three human research domains

Estelle Dumas-Mallet^{1,2}, Katherine S. Button³, Thomas Boraud¹, Francois Gonon¹ and Marcus R. Munafò^{4,5}

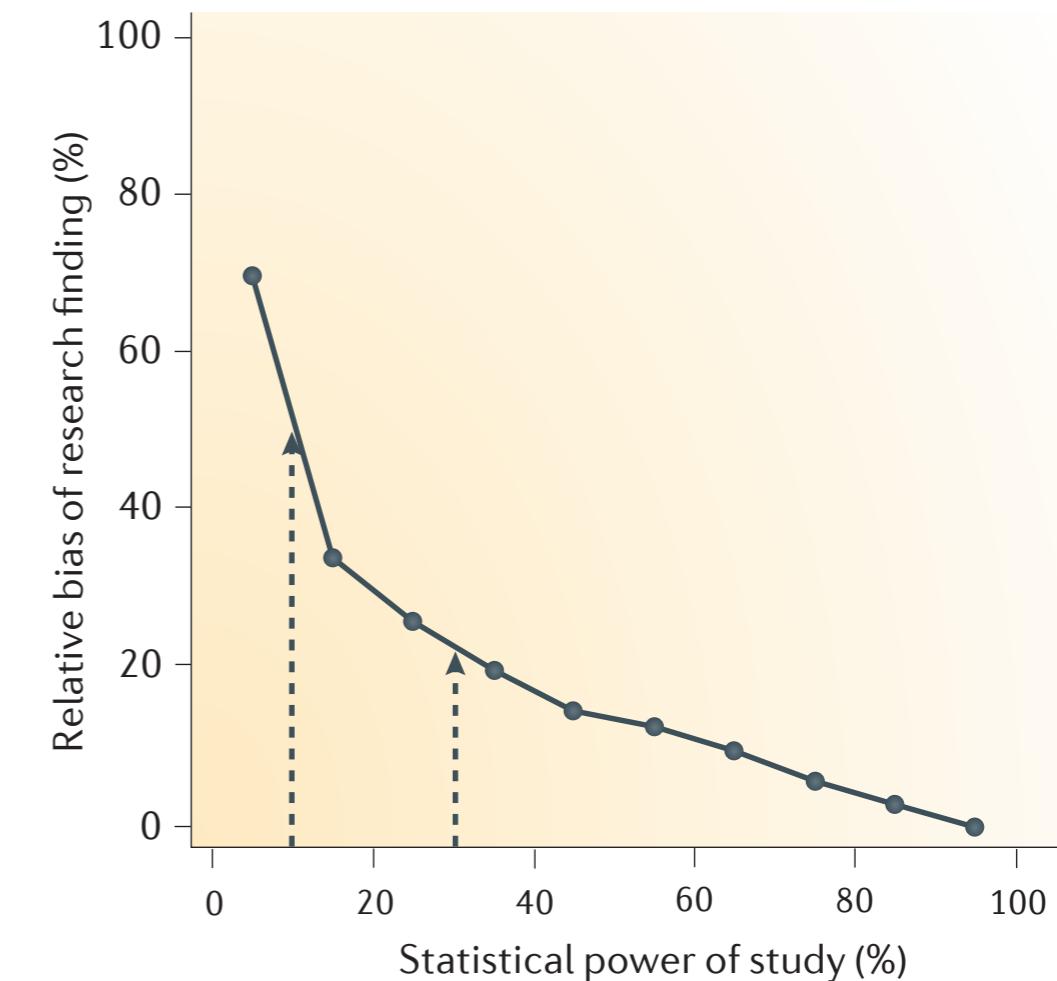
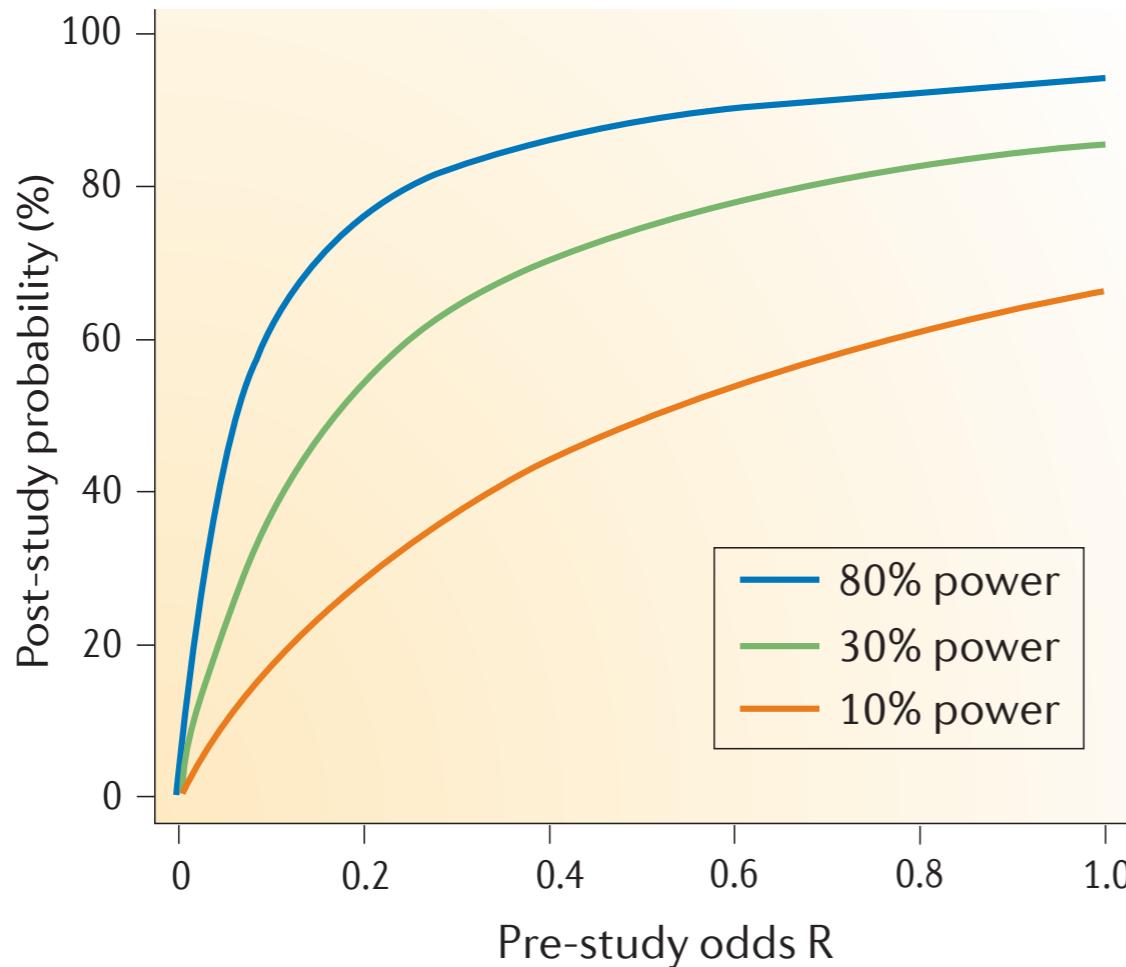


Low power -> unreliable science

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

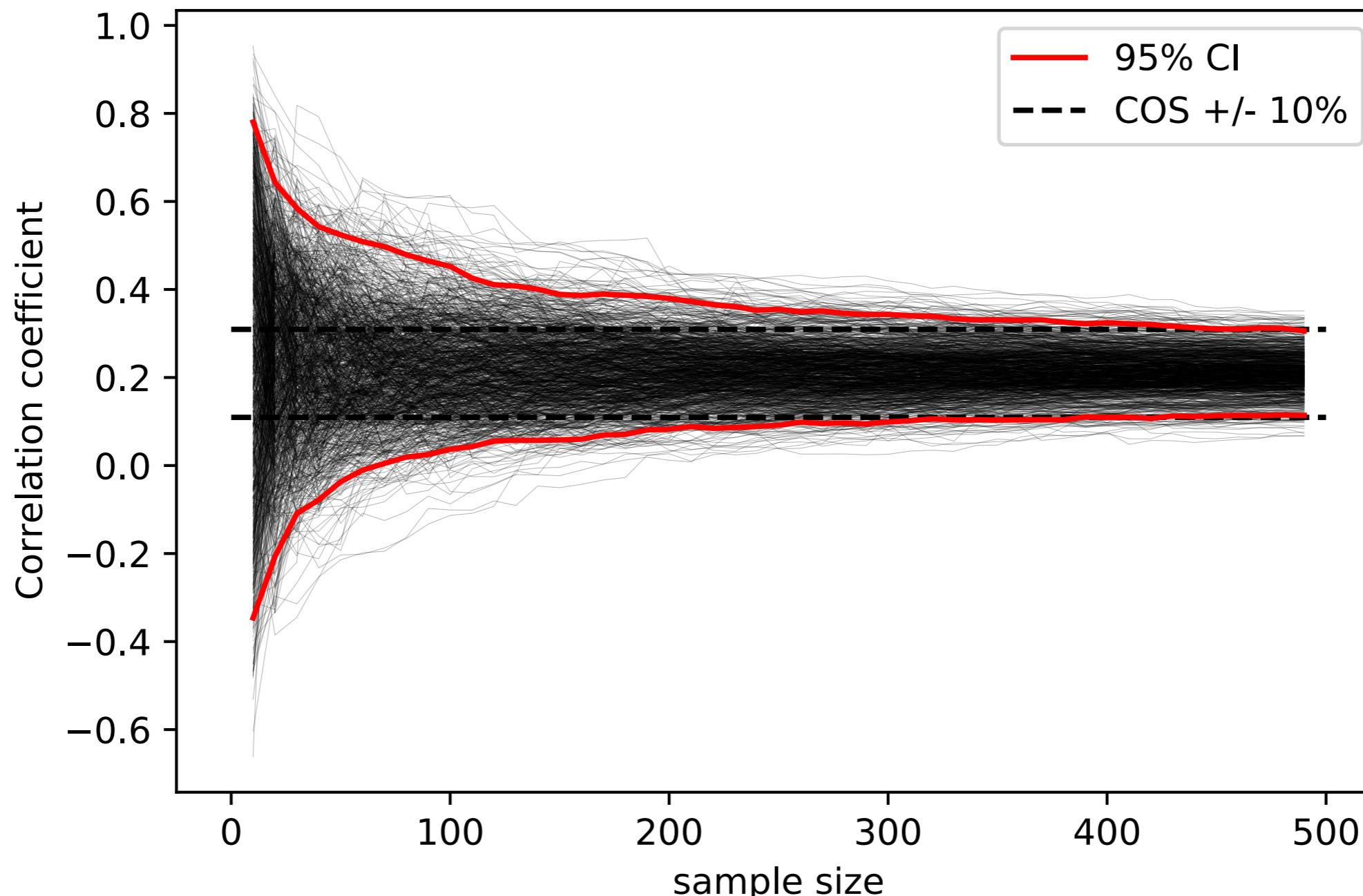
$$\text{PPV} = ([1 - \beta] \times R) / ([1 - \beta] \times R + \alpha)$$

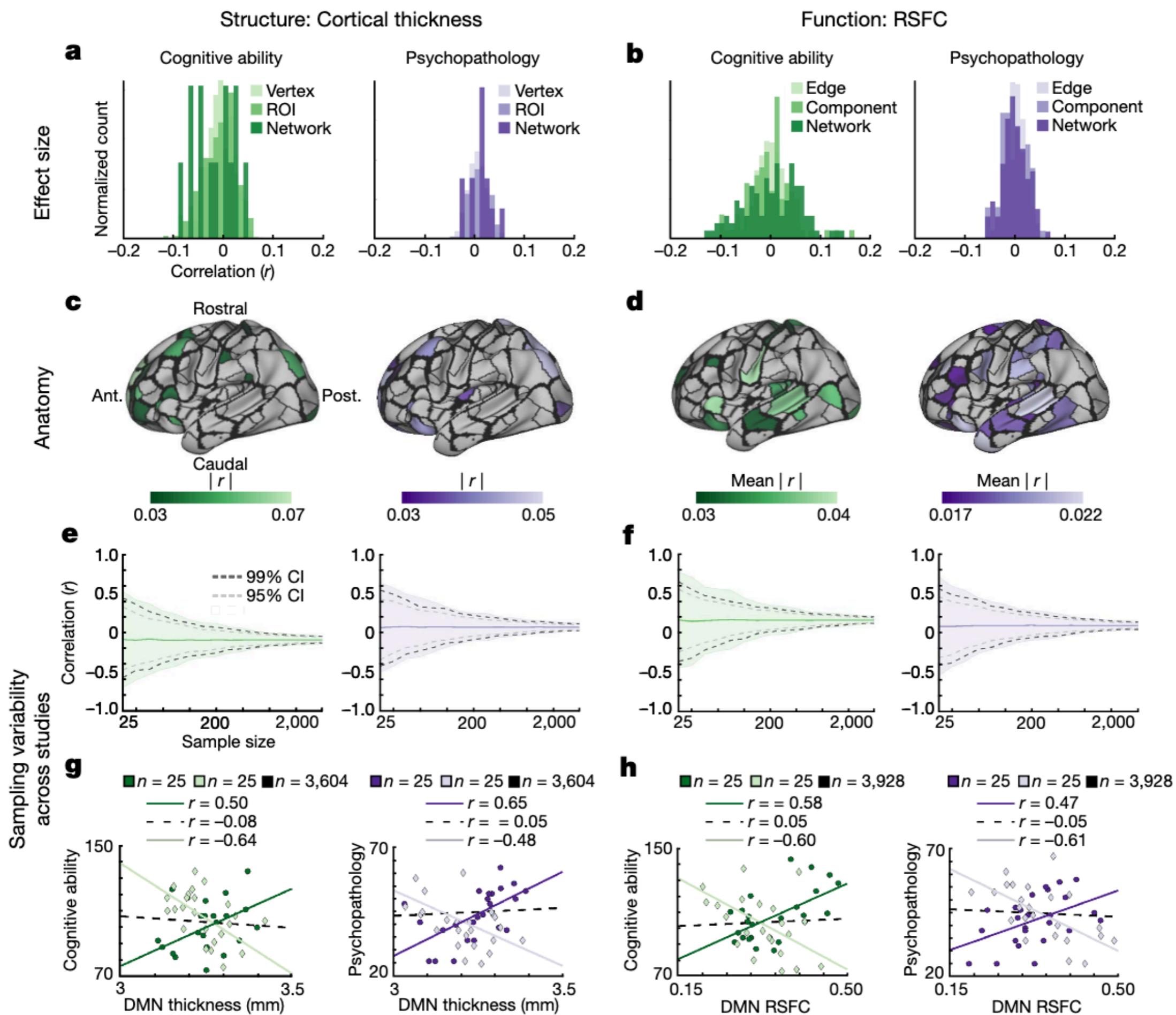
Winner's Curse: overestimation of effect sizes for significant results



Small samples = high instability of statistical estimates

correlation b/w BMI and Hb A1c in NHANES
(full sample $r = 0.21$, $n = 5951$)

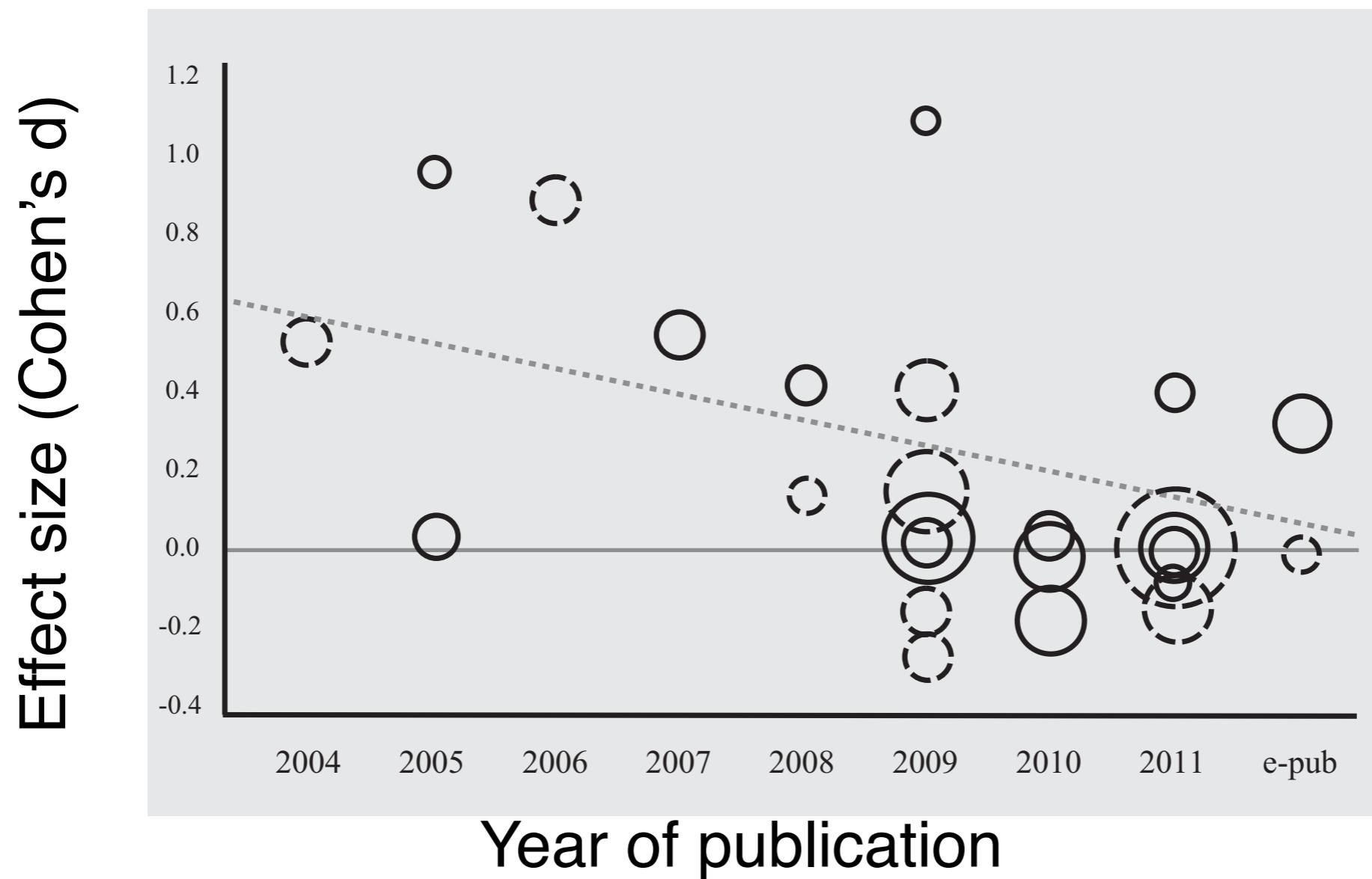




Small samples + publication bias: the case of candidate genes

A Systematic Review and Meta-Analysis on the Association Between BDNF val⁶⁶met 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}



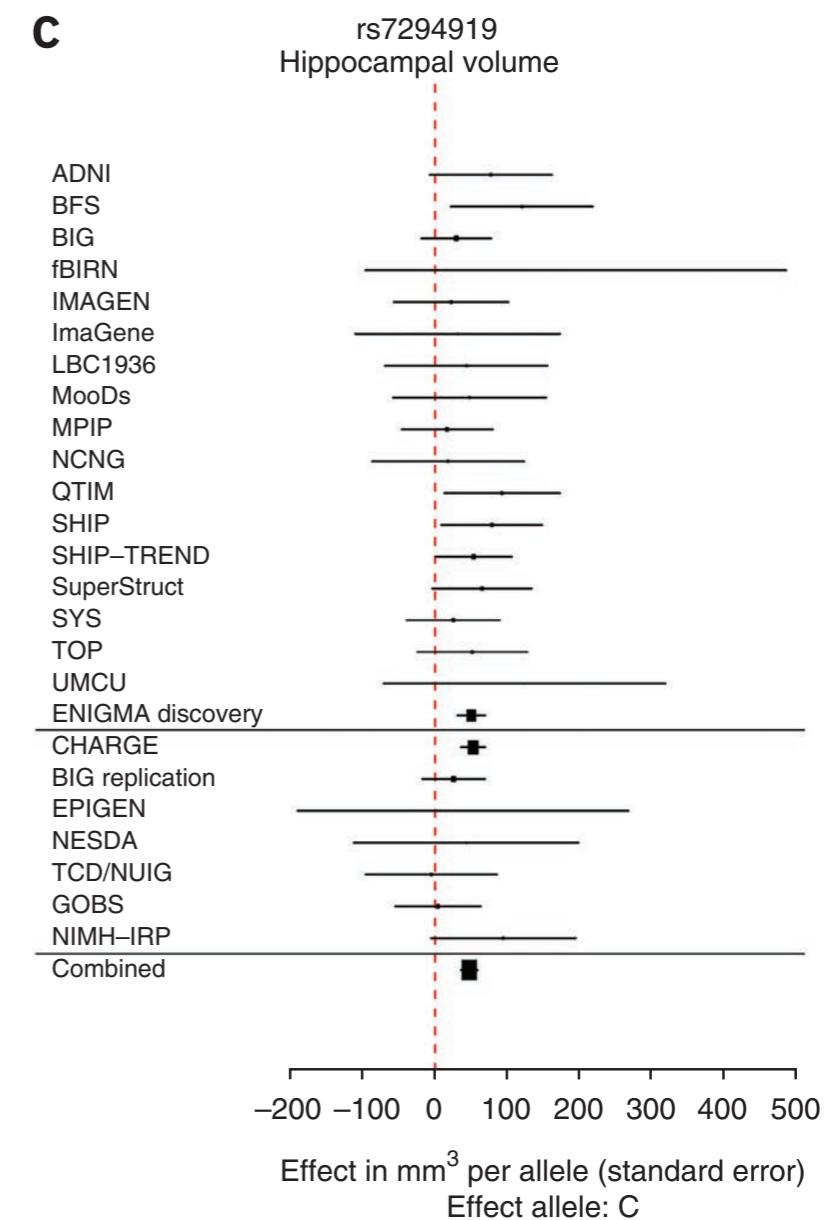
Candidate gene associations fail in well-powered GWAS

Identification of common variants associated with human hippocampal and intracranial volumes

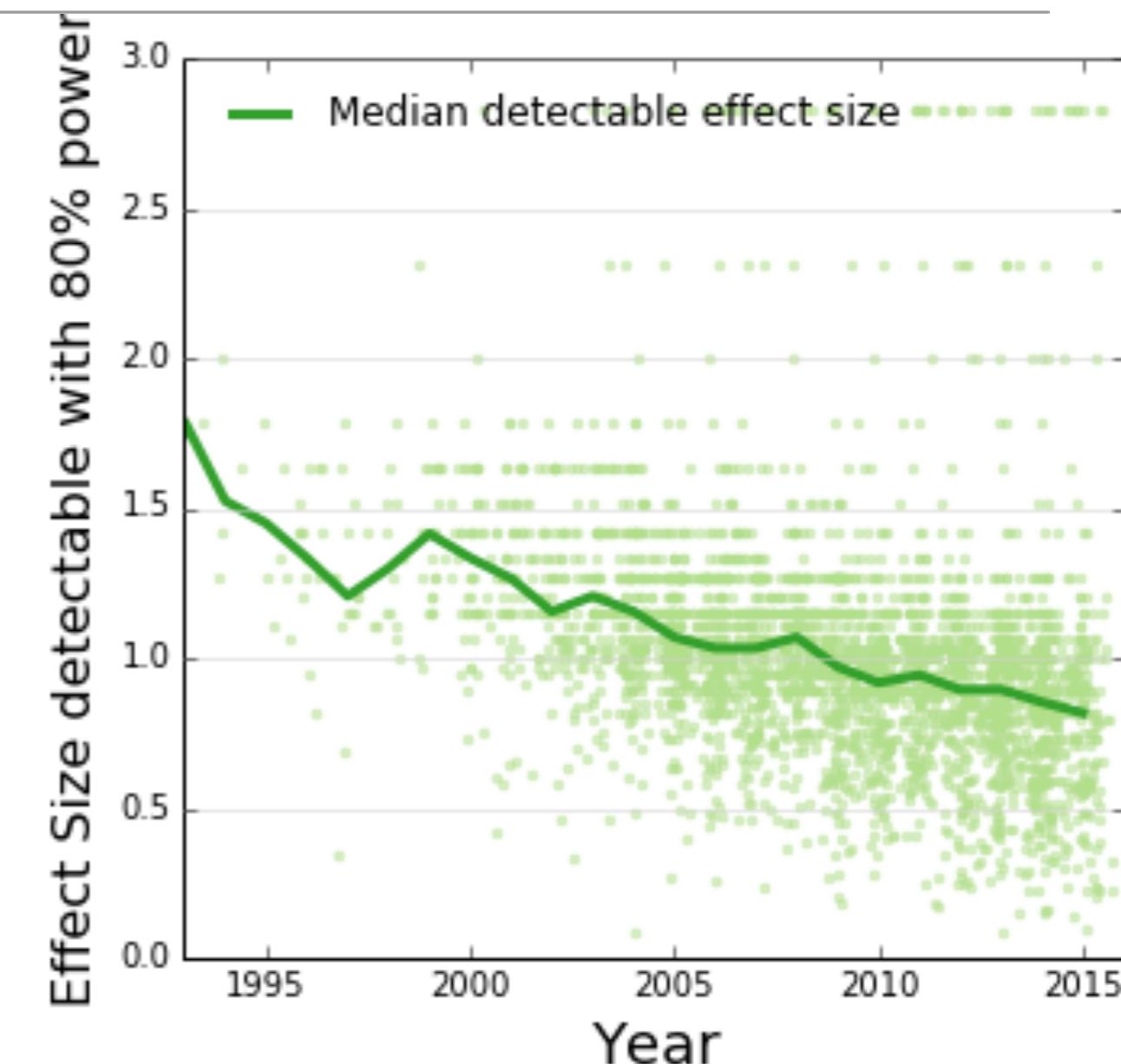
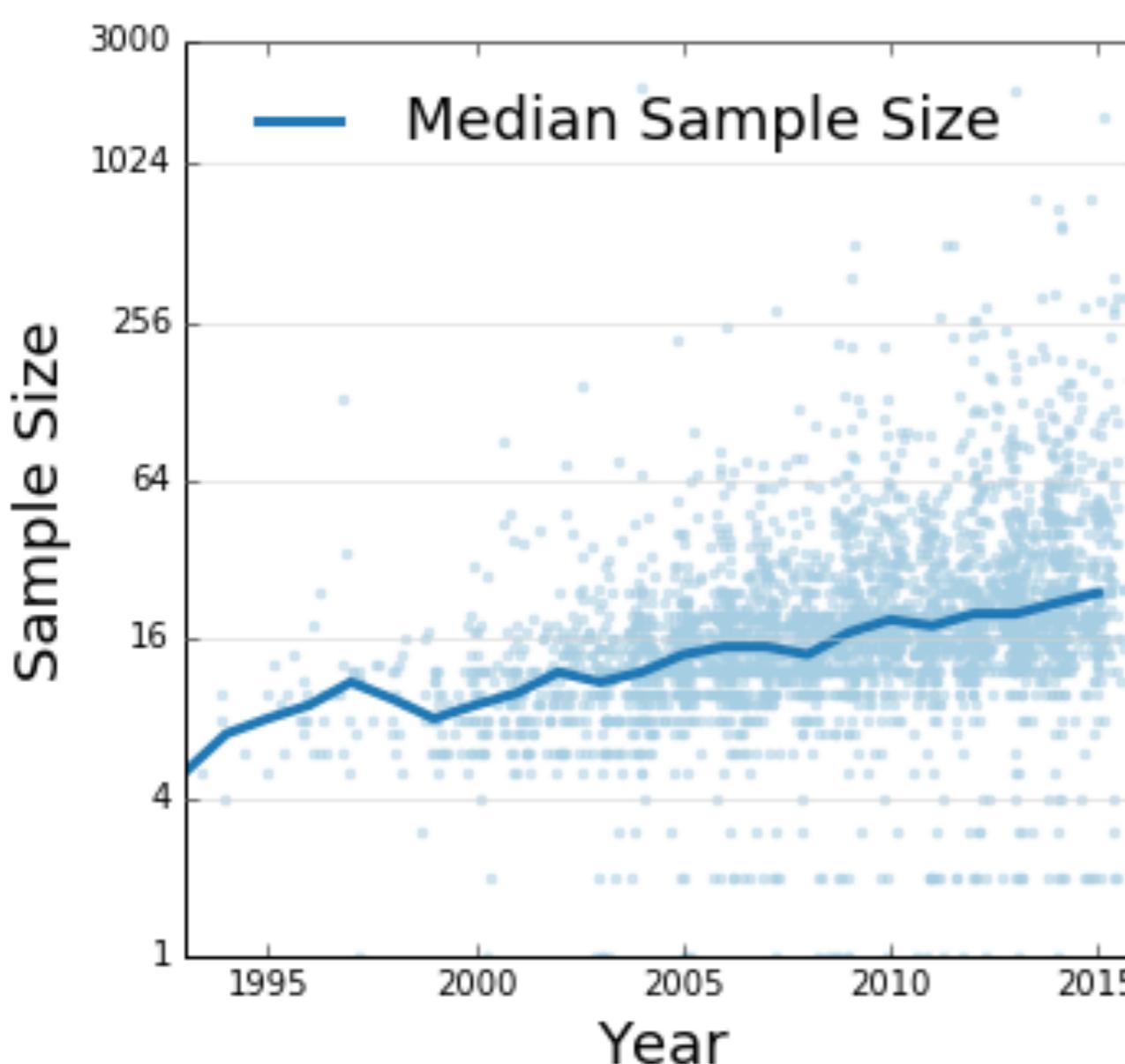
Jason Stein et al. for the Enigma Consortium

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

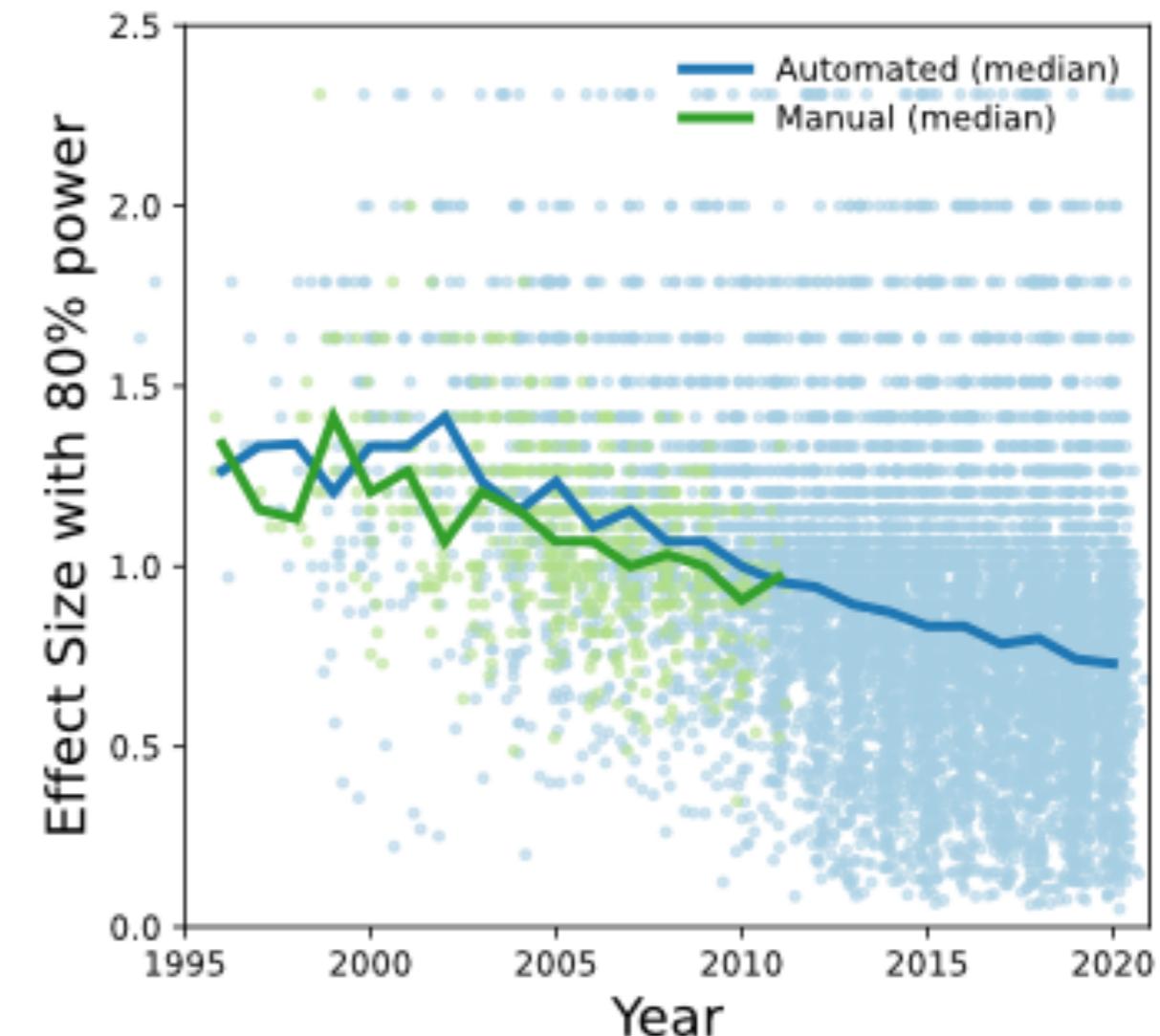
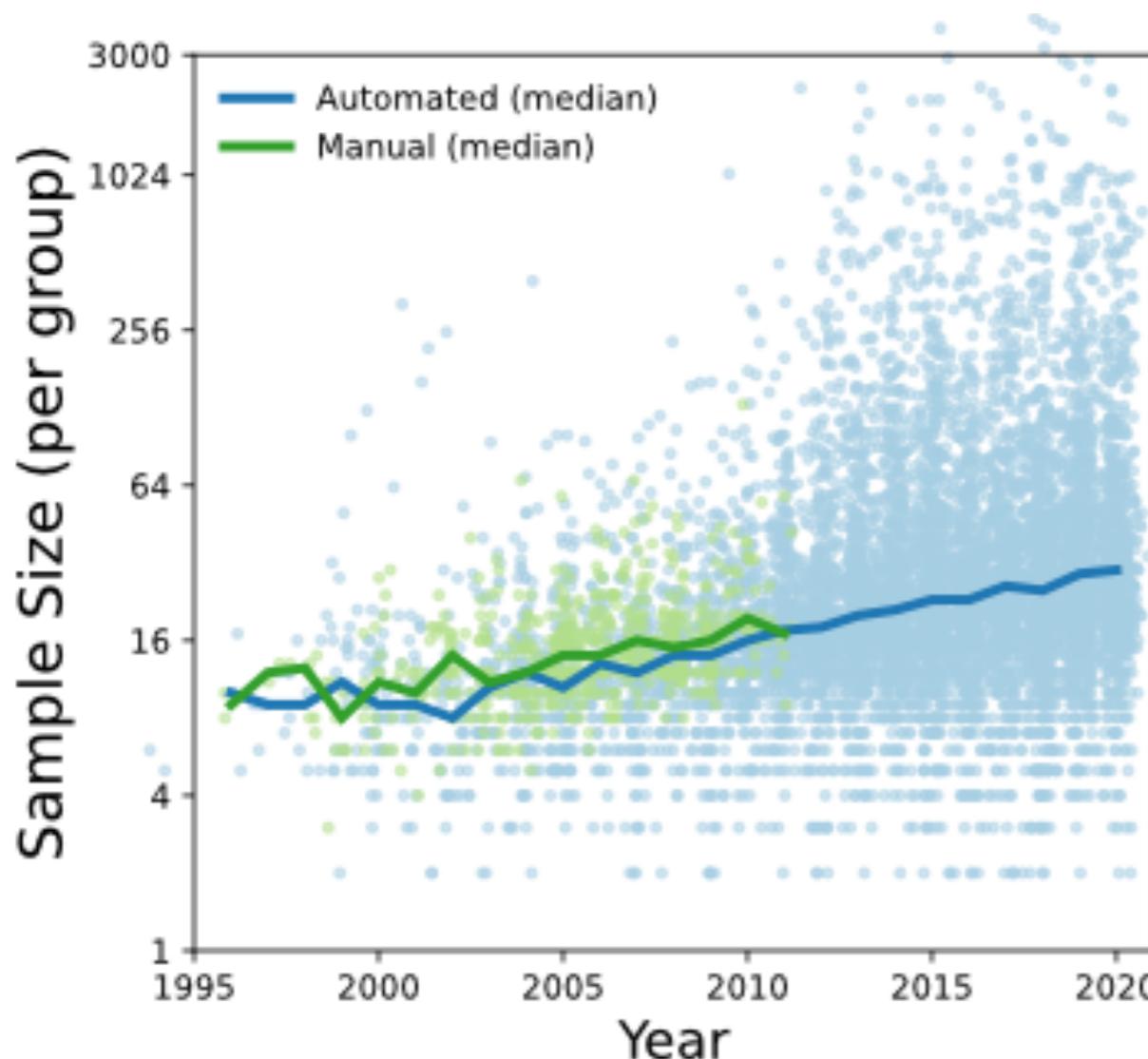


How well powered are fMRI studies?



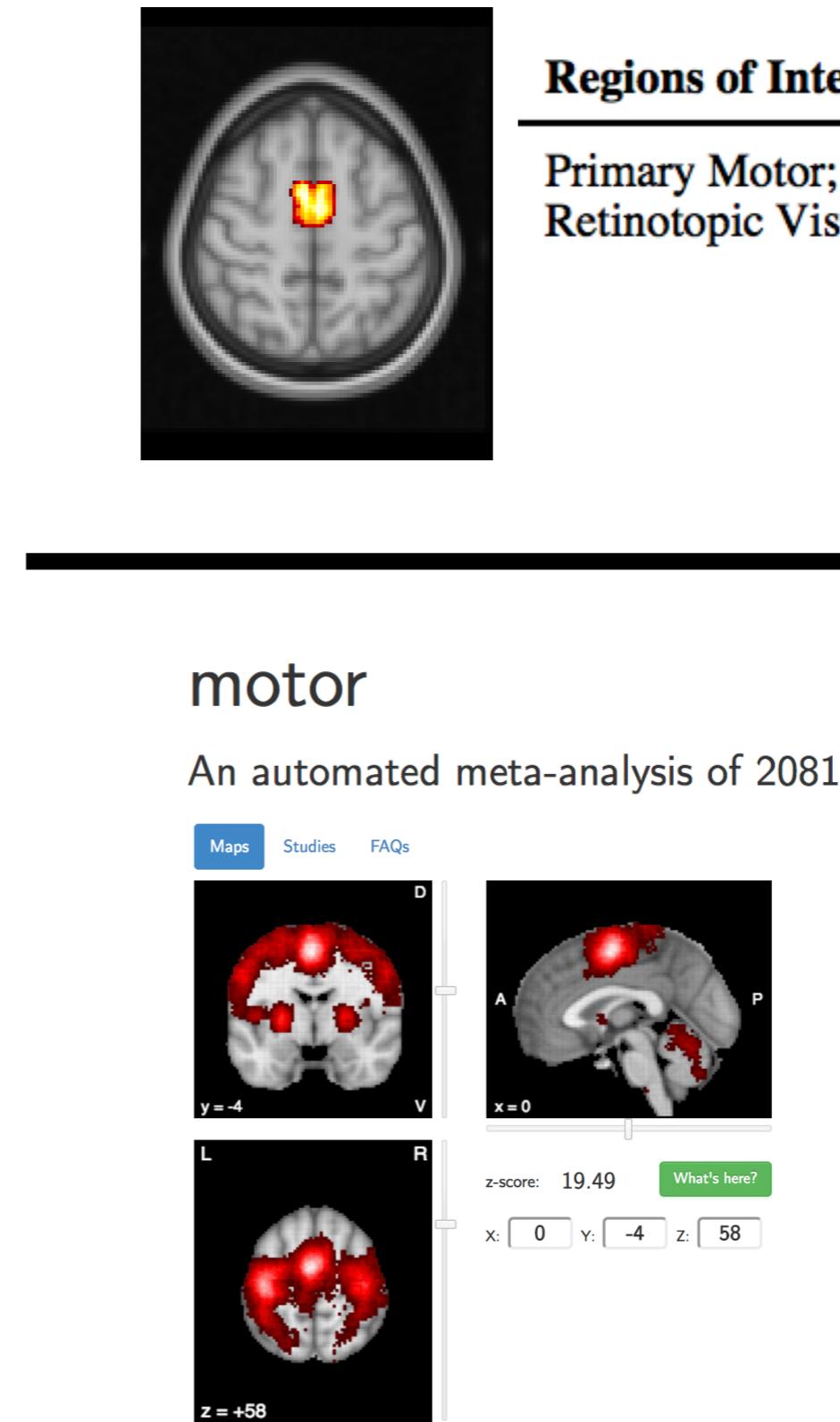
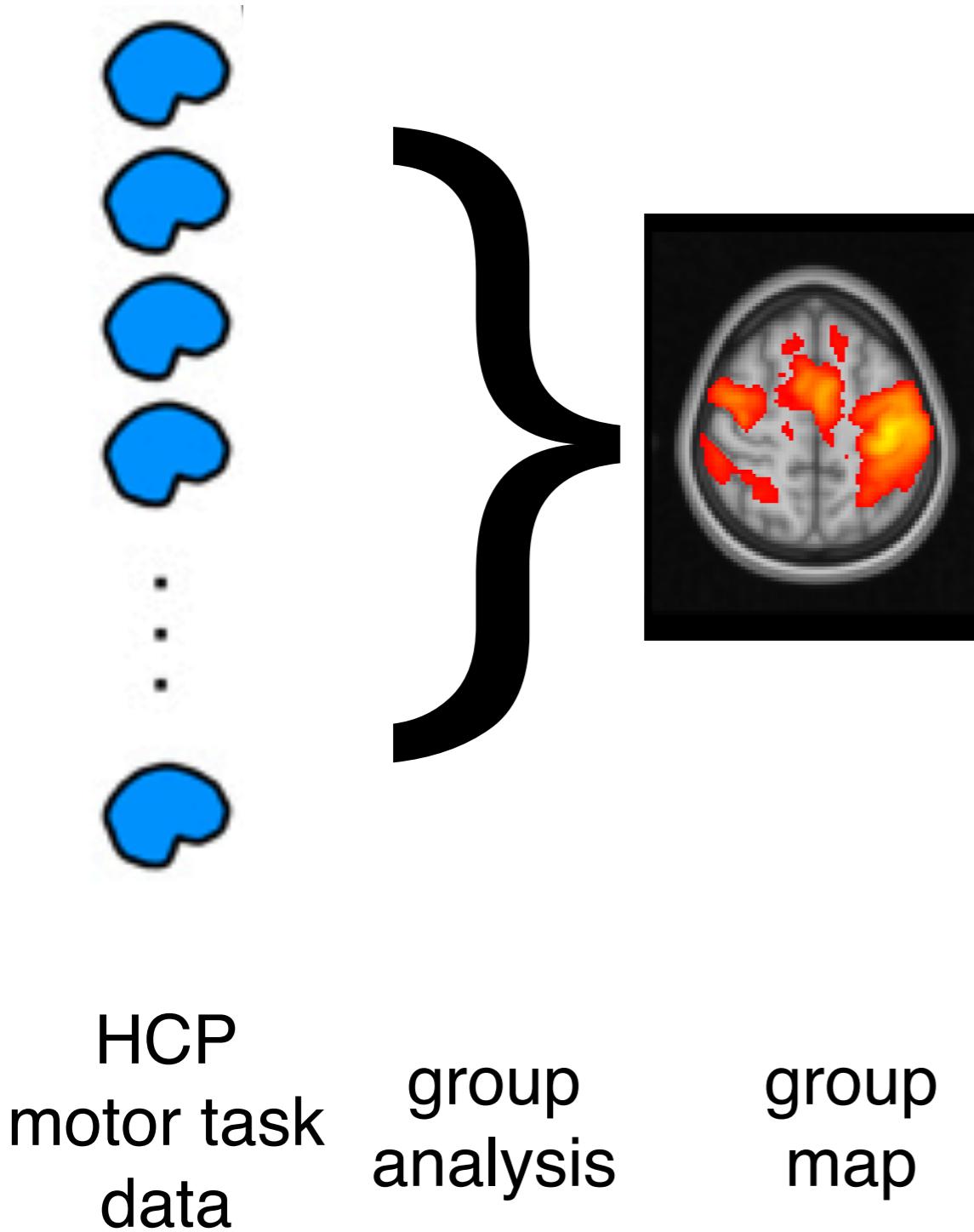
- Median study in 2015 was powered for find a single 200 voxel activation with $d \sim 0.75$
- Is that plausible?

How well powered are fMRI studies?



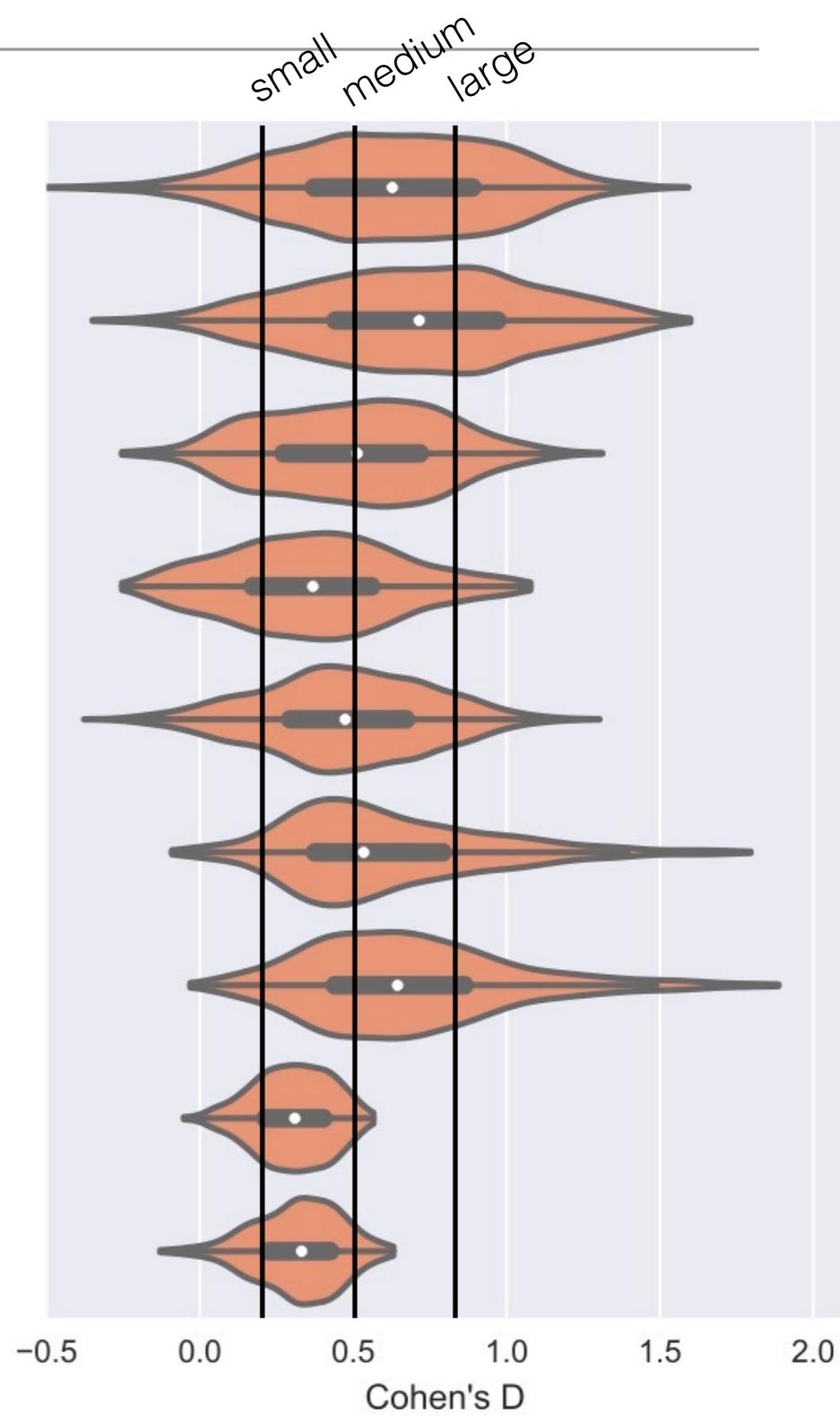
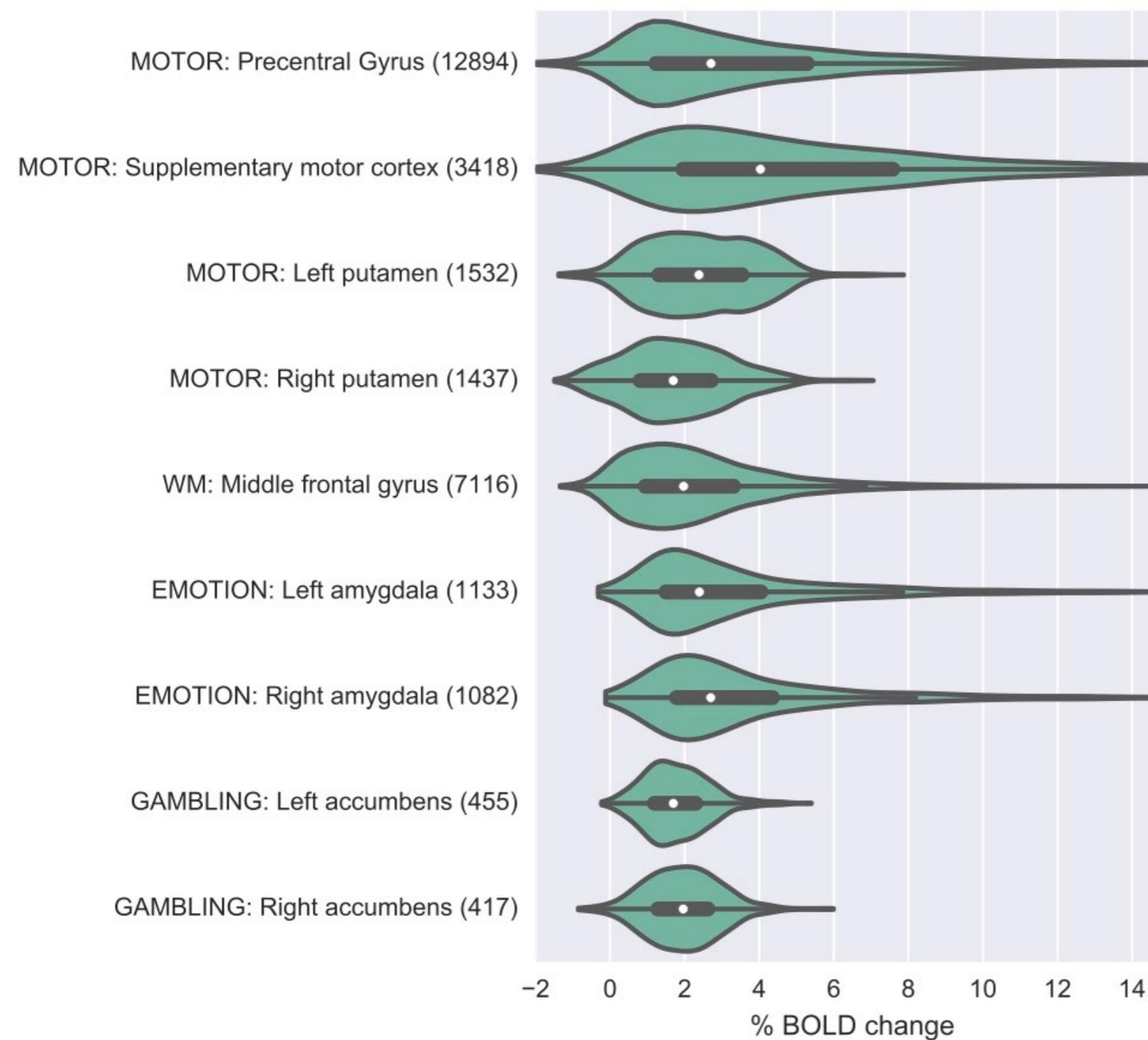
- Median study in 2020 ($n=30/\text{group}$) was powered to find a single 200 voxel activation with $d \sim 0.73$
- Is that plausible?

Estimating realistic effect sizes



unbiased
ES
estimate

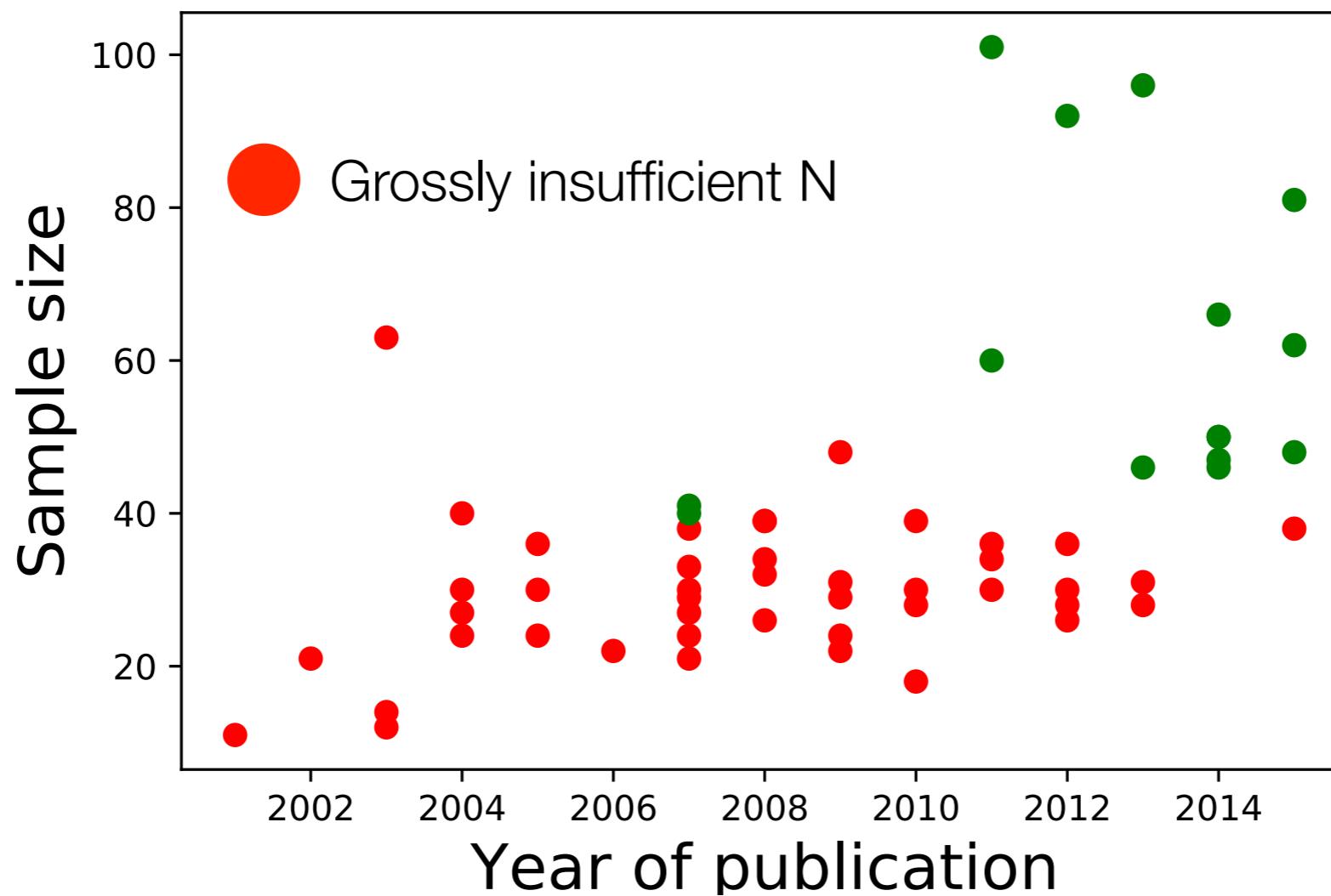
What are realistic effect sizes for fMRI?



Estimated from HCP task data
using combined anatomical + neurosynth ROIs

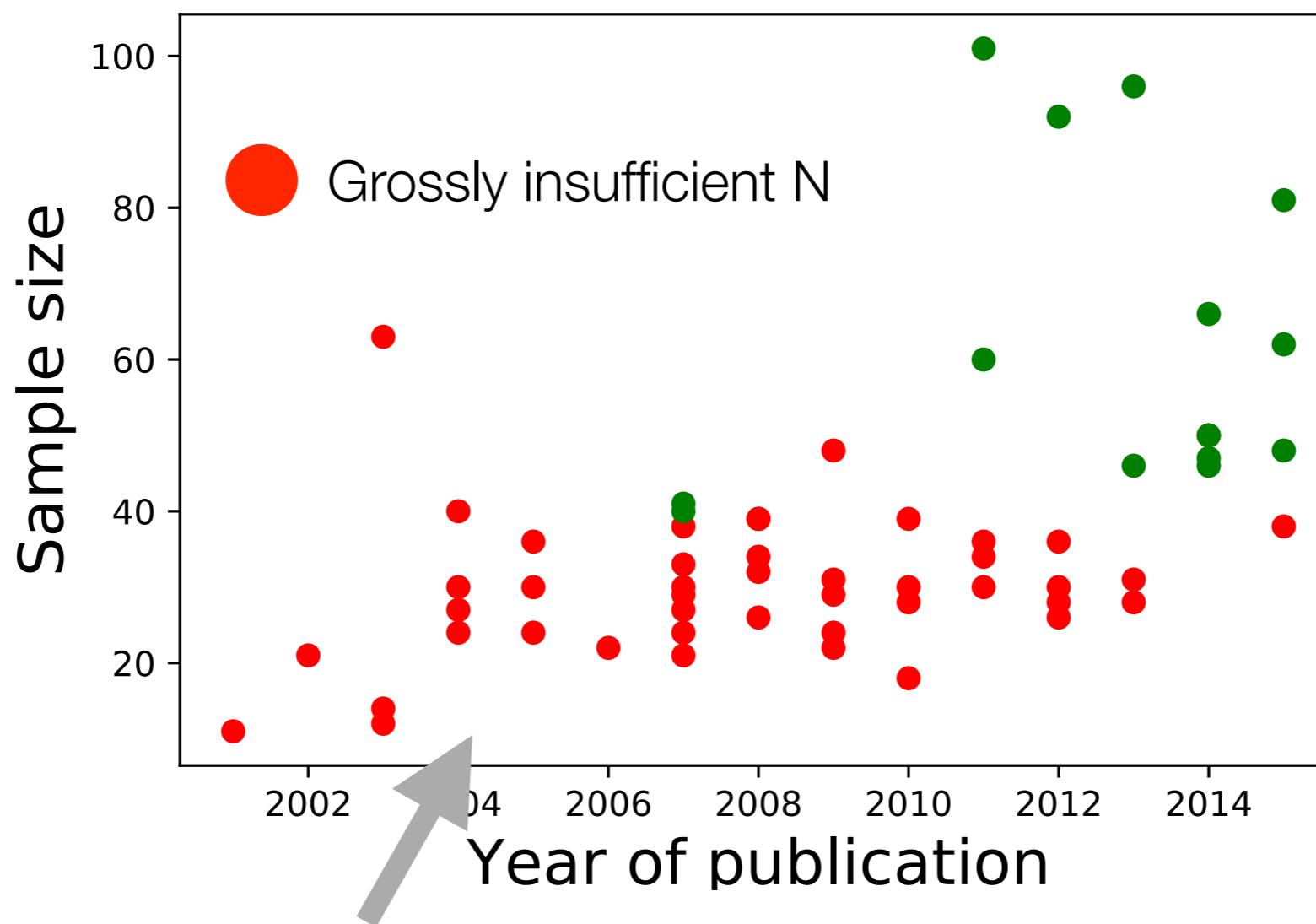
Poldrack et al, 2017, NRN

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 almost certain to generate false results due to insufficient power (and generated about 224 tons of CO₂ in the process - equivalent to about 44 cars driving for a year)

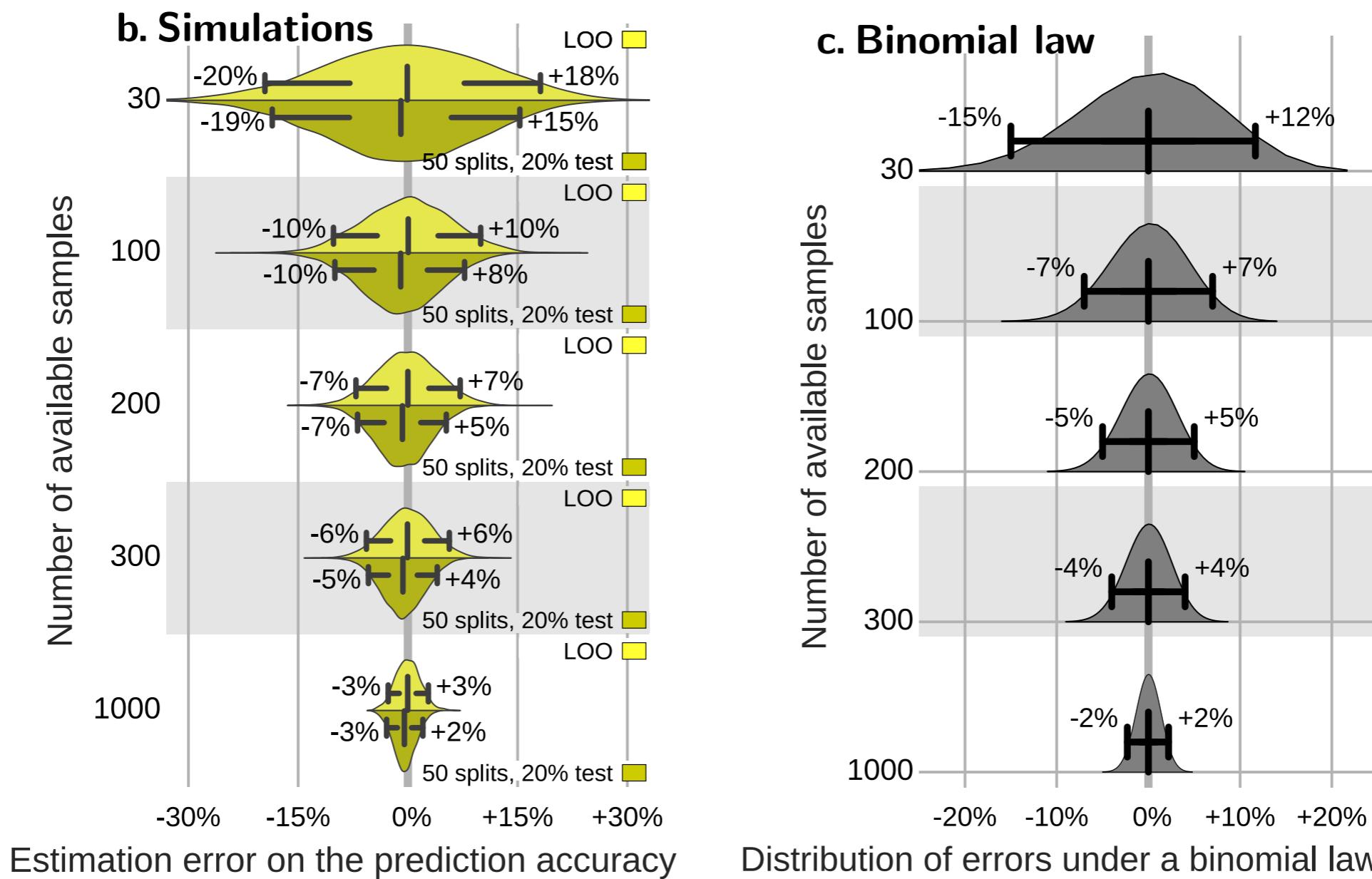
Small samples = variable estimates of predictive accuracy

Cross-validation failure: small sample sizes lead to large error bars

Gaël Varoquaux^{a,b,*}

^a Parietal project-team, INRIA Saclay-île de France, France

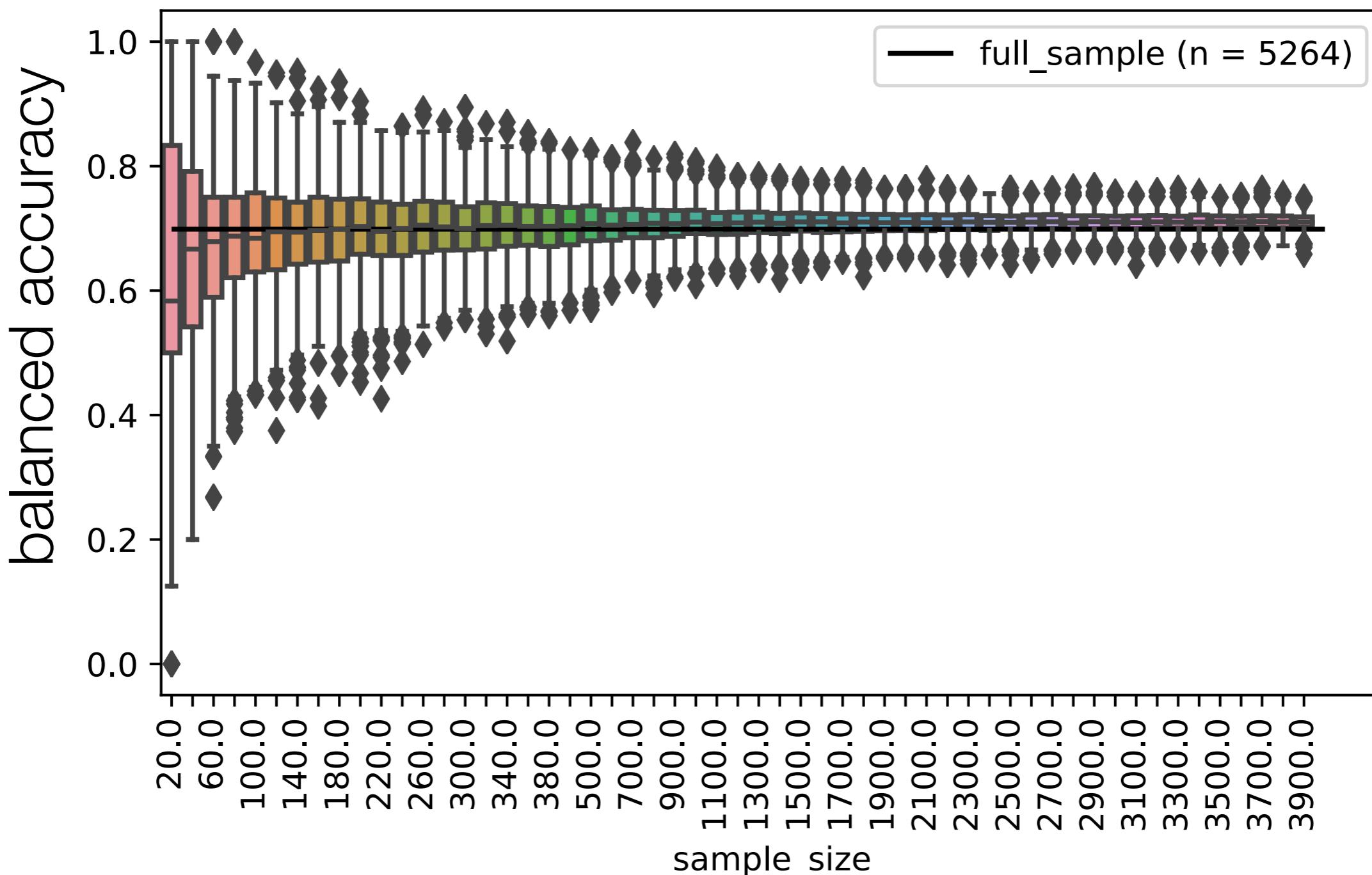
^b CEA/Neurospin bât 145, 91191 Gif-Sur-Yvette, France



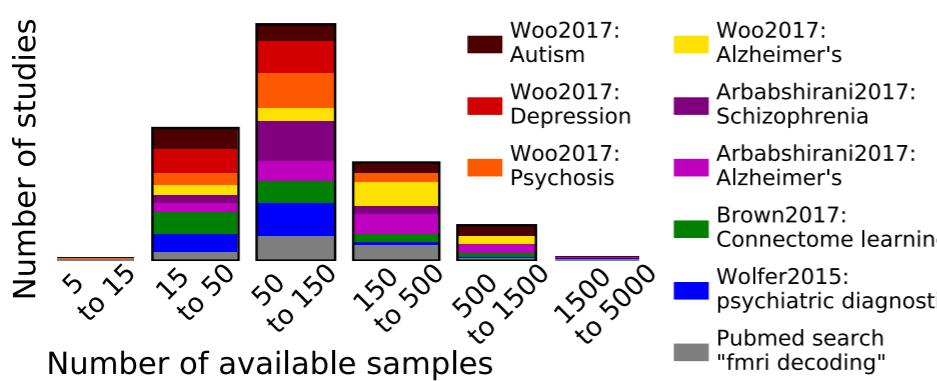
Variability in predictive accuracy on small samples

$$\text{smoker} \sim \text{age} + \text{gender} + \text{bloodcadmium}$$

logistic regression model fit to samples from NHANES



Small samples + publication bias = inflated accuracy estimates

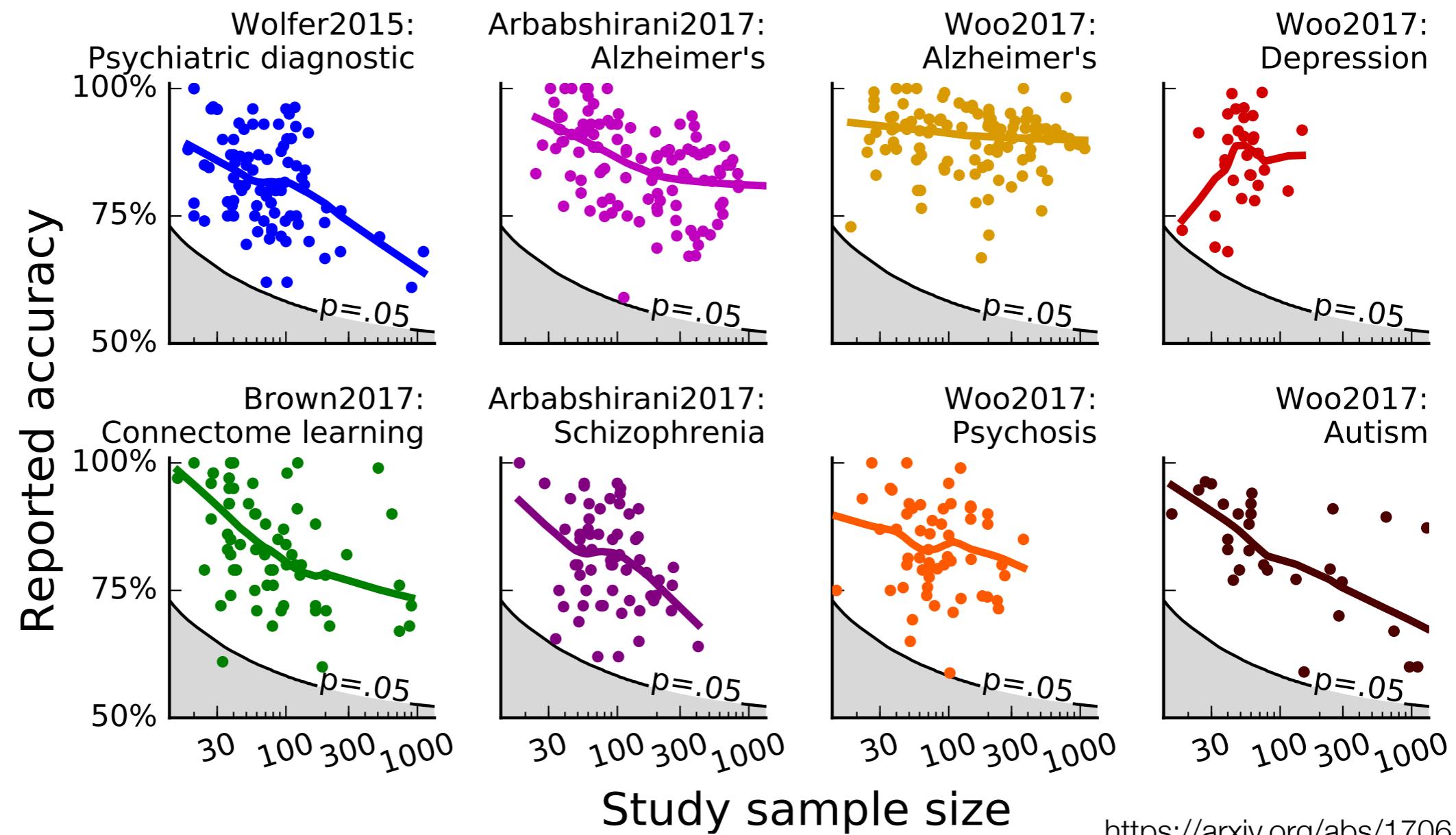


Cross-validation failure: small sample sizes lead to large error bars

Gaël Varoquaux^{a,b,*}

^aParietal project-team, INRIA Saclay-Île de France, France

^bCEA/Neurospin bâti 145, 91191 Gif-Sur-Yvette, France



Doing well-powered science as an early-career researcher

Neuron

NeuroView

The Costs of Reproducibility

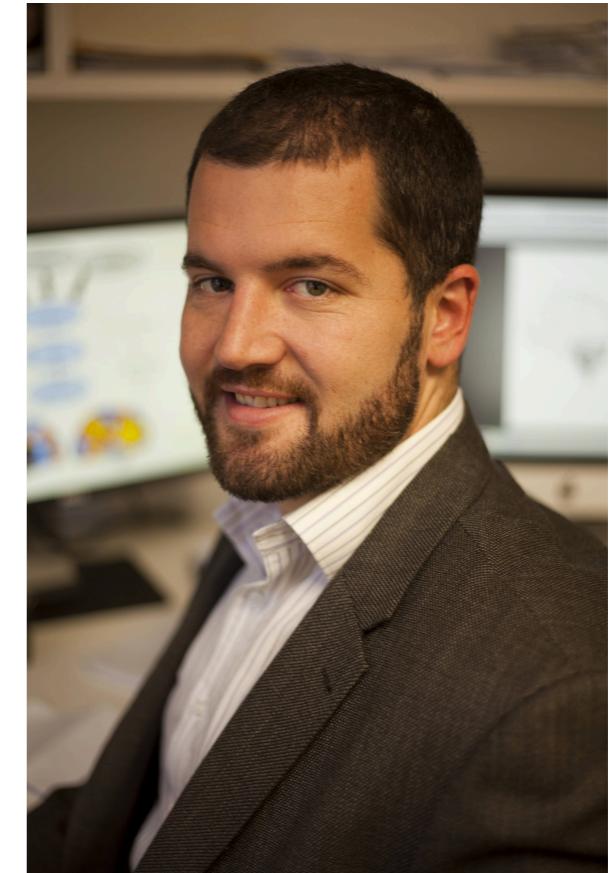
Russell A. Poldrack^{1,*}

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!

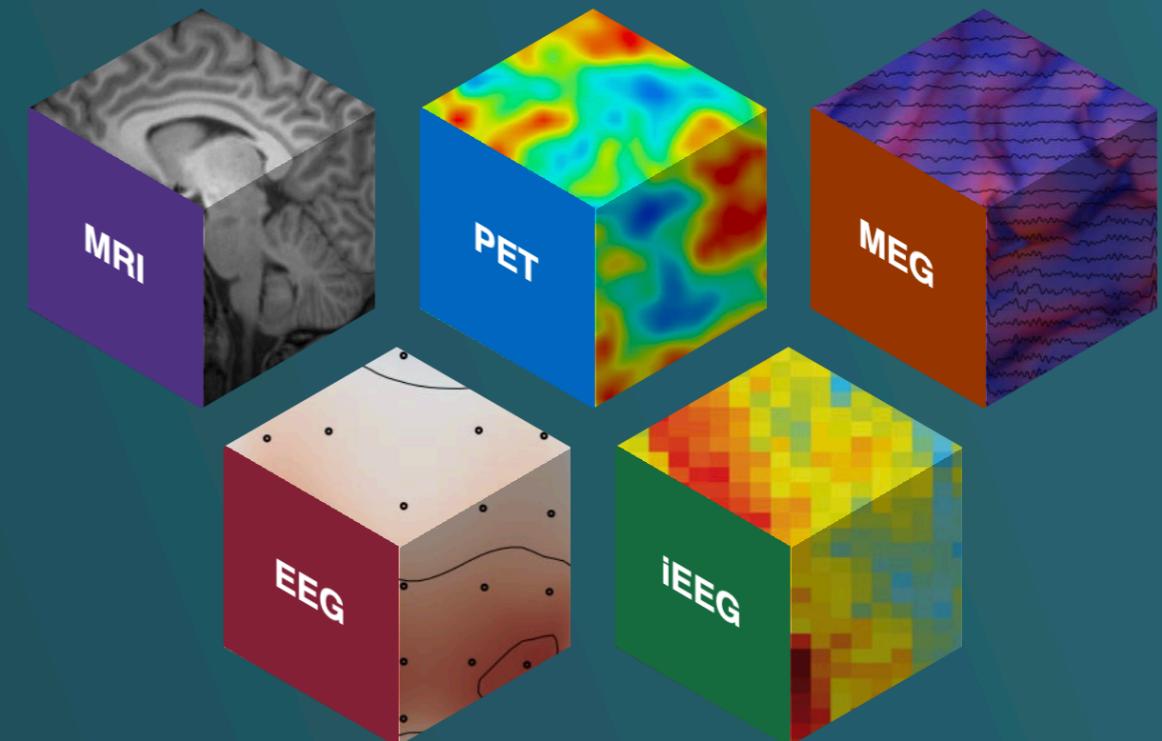




A free and open platform for validating and sharing BIDS-compliant **MRI**, **PET**, **MEG**, **EEG**, and **iEEG** data

27,543 Participants

720 Public Datasets



Browse by Modalities



Or

Search



SIGN IN

G Google

ID ORCID



Validation Using BIDS

The [Brain Imaging Data Structure](#) (BIDS) is an emerging standard for the organization of neuroimaging data.

Want to contribute to BIDS?

Visit the [Google discussion group](#) to contribute.



OpenNeuro Runs on DataLad

Want to access OpenNeuro datasets with DataLad? Visit the [dataset collection on GitHub](#).

A data management solution built on [Git](#) and [Git-annex](#). Read more about [DataLad](#)

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis



PLoS Medicine | www.plosmedicine.org

0696

August 2005 | Volume 2 | Issue 8 | e124

- 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.

Processing step	Reason	Options [suboptions]	Number of plausible options
Motion correction	Correct for head motion during scanning	• ‘Interpolation’ [linear or sinc] • ‘Reference volume’ [single or mean]	4
Slice timing correction	Correct for differences in acquisition timing of different slices	‘No’, ‘before motion correction’ or ‘after motion correction’	3
Field map correction	Correct for distortion owing to magnetic susceptibility	‘Yes’ or ‘no’	2
Spatial smoothing	Increase SNR for larger activations and ensure assumptions of GRF theory	‘FWHM’ [4 mm, 6 mm or 8 mm]	3
Spatial normalization	Warps an individual brain to match a group template	‘Method’ [linear or nonlinear]	2
High-pass filter	Remove low-frequency nuisance signals from data	‘Frequency cut-off’ [100 s or 120 s]	2
Head motion regressors	Remove remaining signals owing to head motion via statistical model	‘Yes’ or ‘no’ [if yes: 6/12/24 parameters or single time point ‘scrubbing’ regressors]	5
Haemodynamic response	Account for delayed nature of haemodynamic response to neuronal activity	• ‘Basis function’ [‘single-gamma’ or ‘double-gamma’] • ‘Derivatives’ [‘none’, ‘shift’ or ‘dispersion’]	6
Temporal autocorrelation model	Model for the temporal autocorrelation inherent in fMRI signals	‘Yes’ or ‘no’	2
Multiple-comparison correction	Correct for large number of comparisons across the brain	‘Voxel-based GRF’, ‘cluster-based GRF’, ‘FDR’ or ‘non-parametric’	4
Total possible workflows			69,120

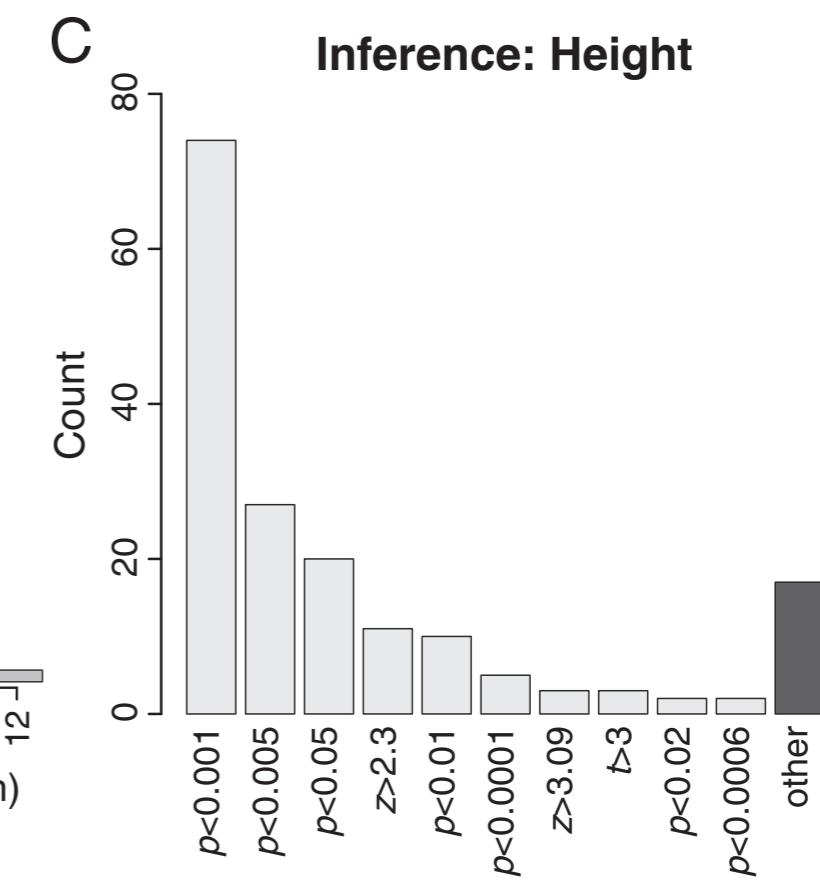
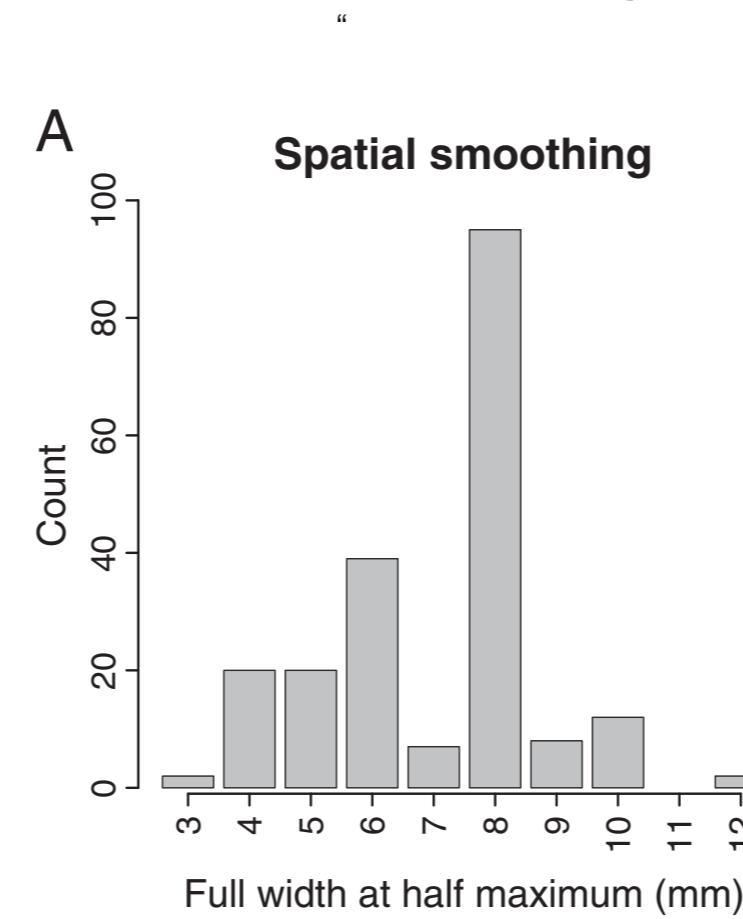
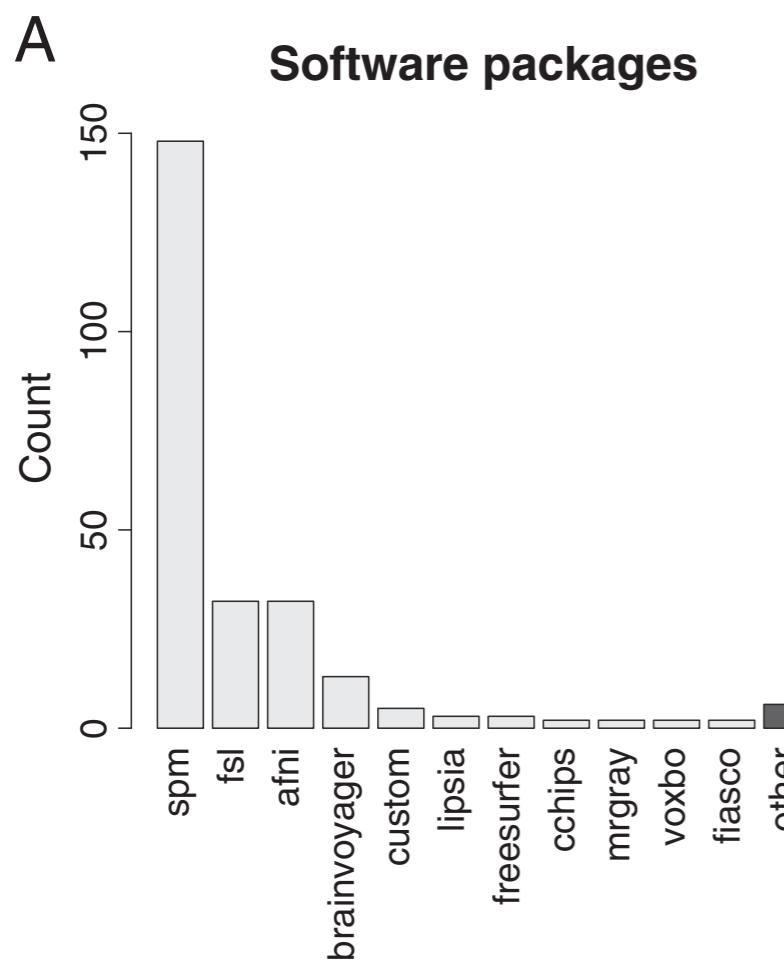
FDR, false discovery rate; FWHM, full width at half maximum; GRF, Gaussian random field; SNR, signal-to-noise ratio.

The secret lives of experiments: Methods reporting in the fMRI literature

Joshua Carp

NeuroImage 63 (2012) 289–300

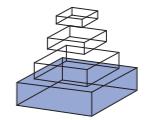
“data collection and analysis methods were highly flexible across studies, with nearly as many unique analysis pipelines as there were studies in the sample [241].”



Threats to reproducibility: Methodological flexibility

frontiers in
NEUROSCIENCE

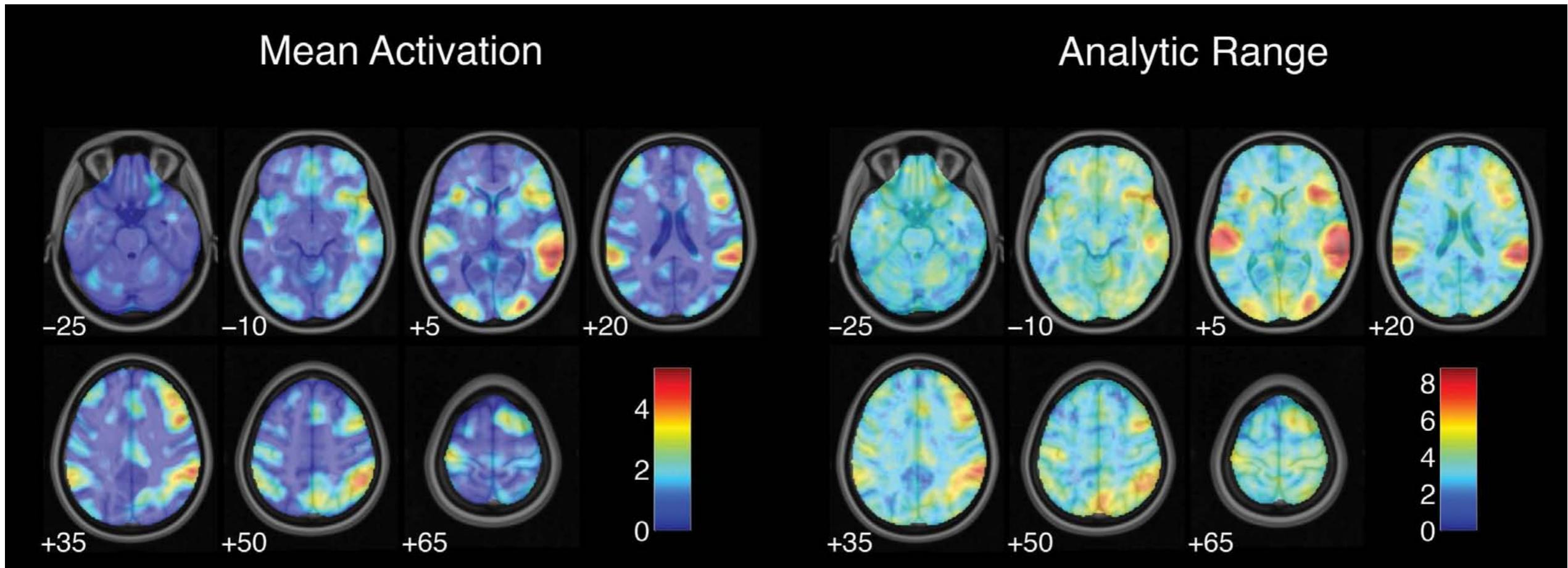
ORIGINAL RESEARCH ARTICLE
published: 11 October 2012
doi: 10.3389/fnins.2012.00149



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

Joshua Carp*

6,912 pipelines

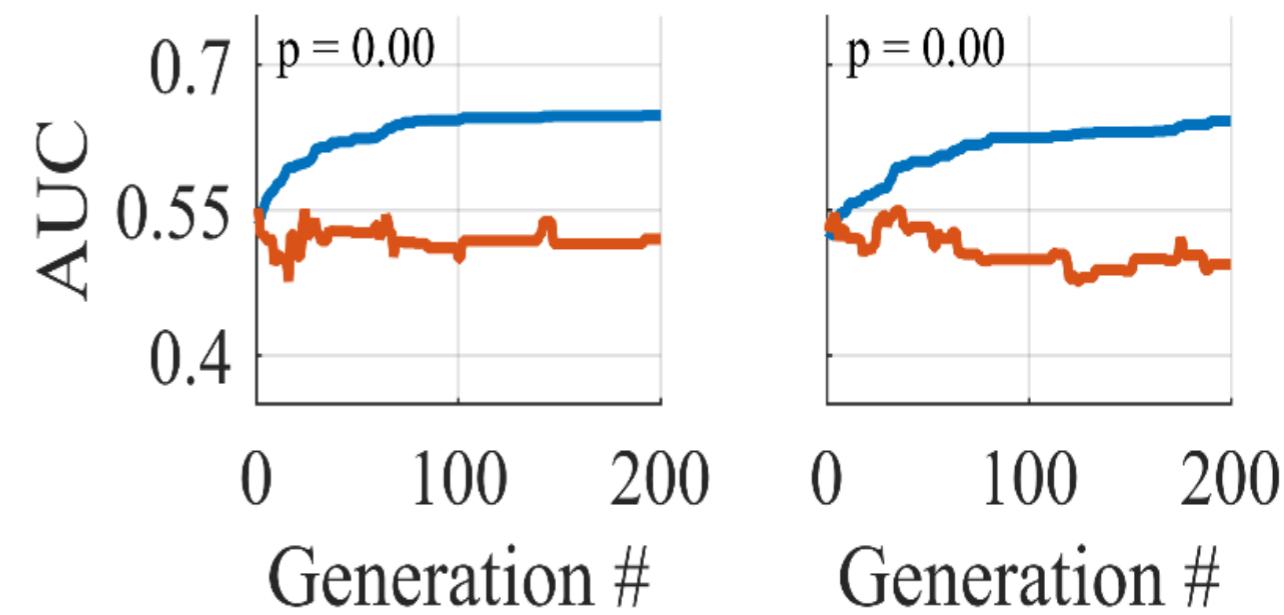


Machine learning can make it worse

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

MICHAEL SKOCIK¹, JOHN COLLINS², CHLOE CALLAHAN-FLINTOFT³, HOWARD
BOWMAN⁴, AND BRAD WYBLE³

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.

Copyright © 2016 Society for Psychophysiological Research

DOI: 10.1111/psyp.12639

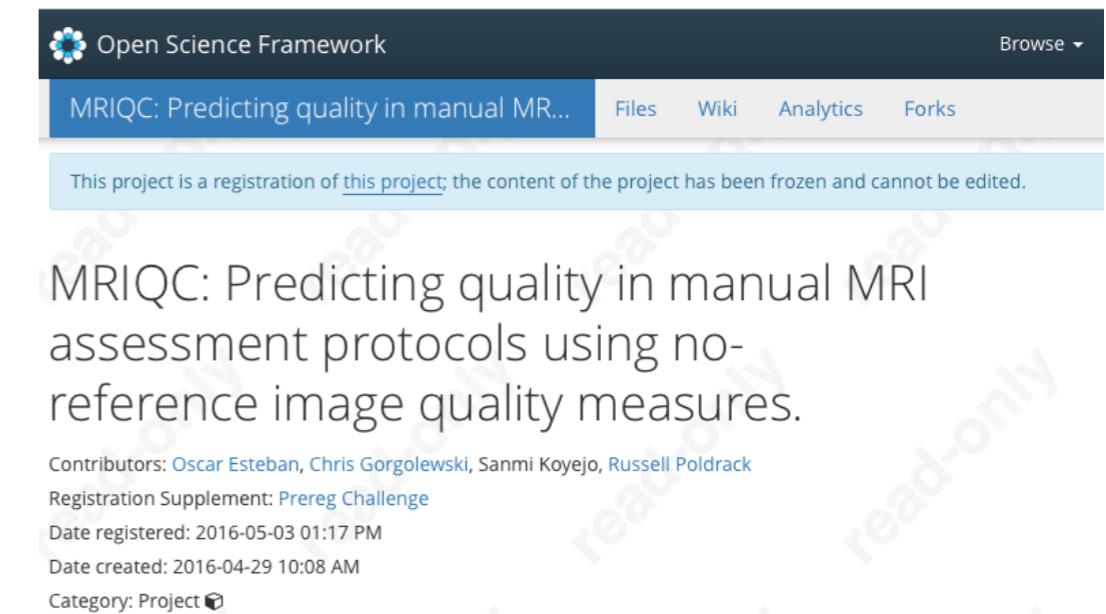
How to get statistically significant effects in any ERP experiment (and why you shouldn't)

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 analysis plan prior to accessing data
 - 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



The screenshot shows a project page on the Open Science Framework. The title is "MRIQC: Predicting quality in manual MR...". The page includes navigation links for "Files", "Wiki", "Analytics", and "Forks". A note at the top states, "This project is a registration of [this project](#); the content of the project has been frozen and cannot be edited." Below this, the project description reads: "MRIQC: Predicting quality in manual MRI assessment protocols using no-reference image quality measures." Contributors listed are Oscar Esteban, Chris Gorgolewski, Sanmi Koyejo, and Russell Poldrack. The registration supplement is identified as "Prereg Challenge". The date registered is 2016-05-03 01:17 PM, and the date created is 2016-04-29 10:08 AM. The category is "Project".

<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

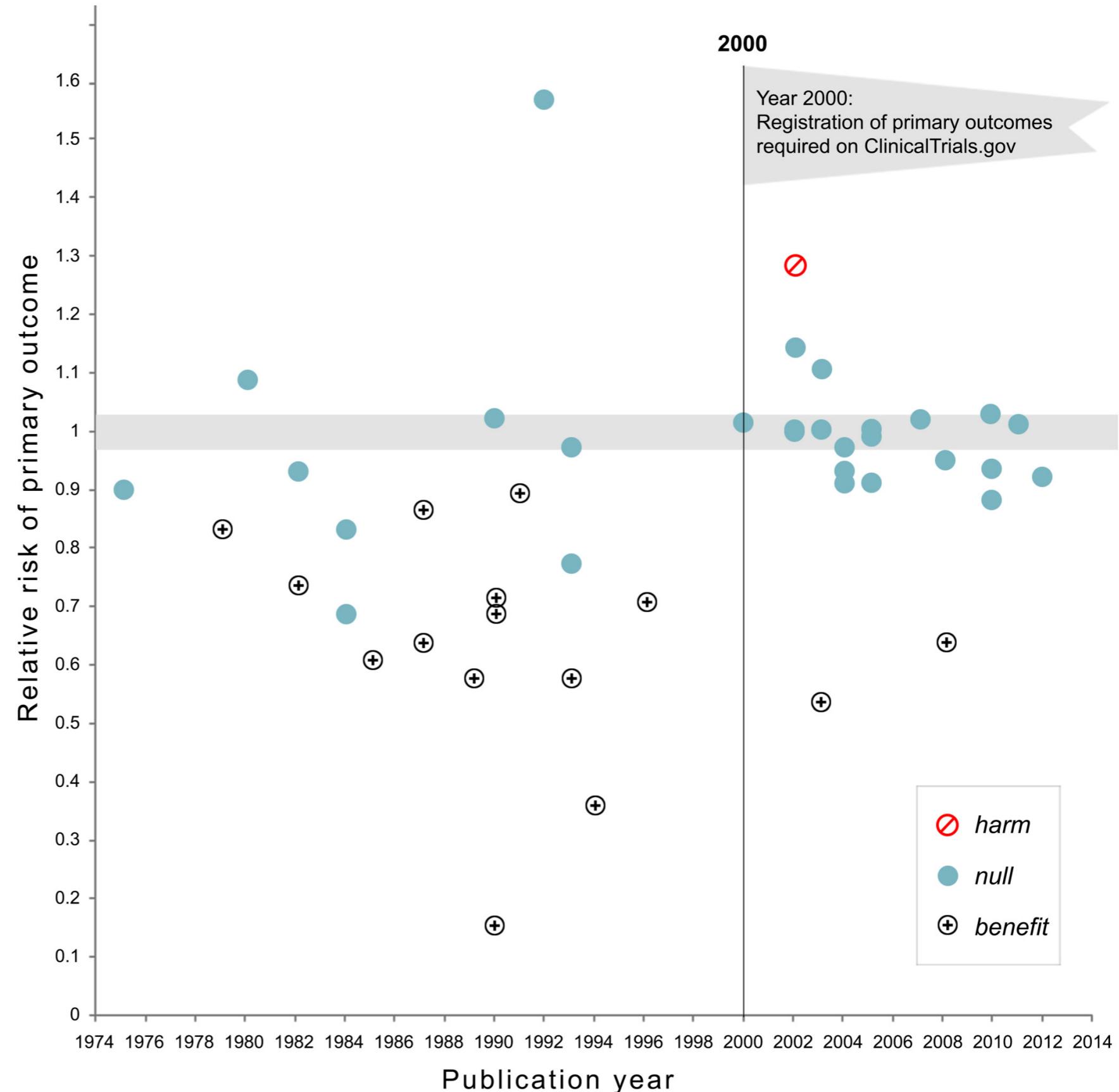


Fig 1. Relative risk of showing benefit or harm of treatment by year of publication for large NHLBI trials on pharmaceutical and dietary supplement interventions. Positive trials are indicated by the plus signs while trials showing harm are indicated by a diagonal line within a circle. Prior to 2000 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.

Pre-registration prevents p-hacking but does
not eliminate analytic variability

How variable are neuroimaging analysis
workflows in the wild?

What is the effect on scientific inferences?

#narps: Neuroimaging Analysis Replication and Prediction Study

Rotem Botvinik-Nezer



Felix Holzmeister



Jeanette Mumford



Anna Dreber



Magnus Johannesson



Tom Schonberg



Tom Nichols



Michael Kirchler



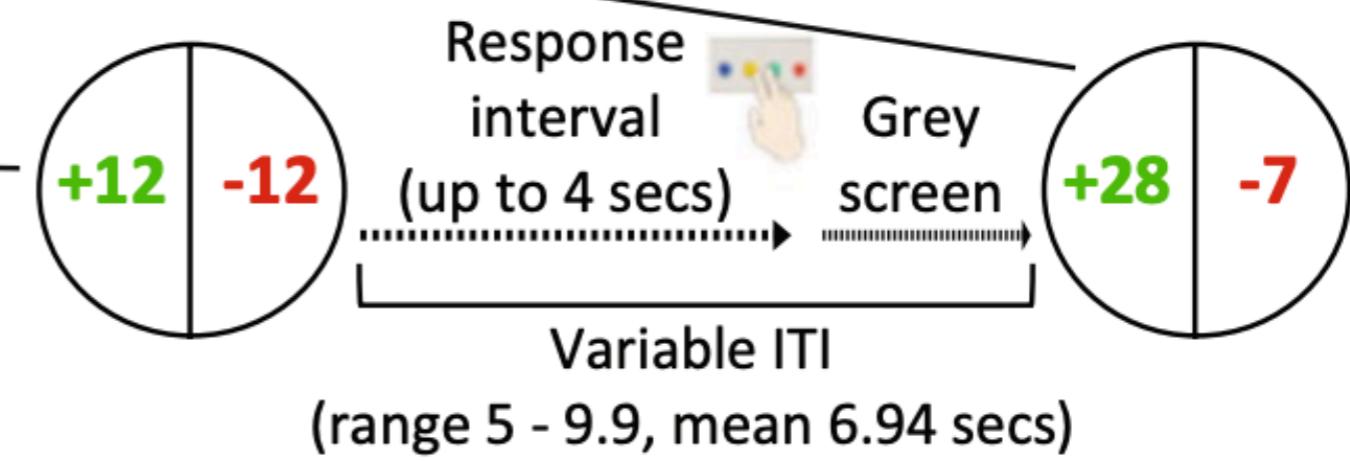
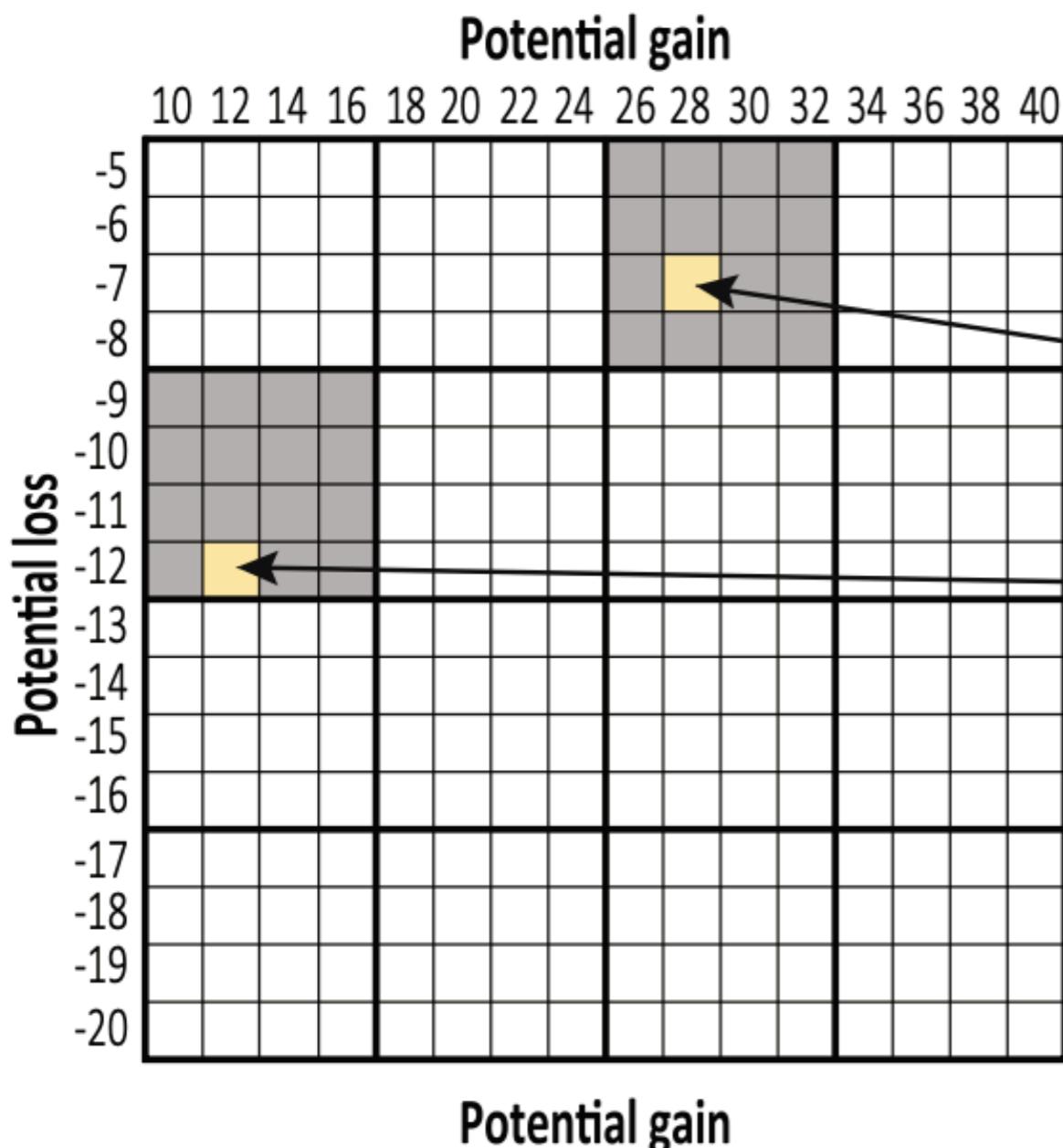
Colin Camerer



Botvinik-Nezer et al., 2020, *Nature*

Data collection: Mixed gambles task

- Functional MRI data for 108 subjects collected at Tel Aviv University



Study recruitment

- Analysis teams recruited via social media and conferences
 - Up to 3 people each
 - Signed non-disclosure agreement
 - Offered co-authorship on initial paper
- Raw dataset (~400GB) distributed to 82 teams via Globus
- Teams given ~3 months to provide Y/N answers to 9 hypothesis tests:
 - Is there significant activation in <brain area> for <feature of experimental design>?
- Using their standard analysis workflow
 - Only restriction was that analysis should include whole brain



Results collection

- Received final results from 70 teams
- Teams reported their yes/no decisions for each hypothesis
 - Also uploaded the whole-brain statistical maps before and after thresholding

How variable are workflows in the wild?

- Teams provided a detailed written description of analysis workflows

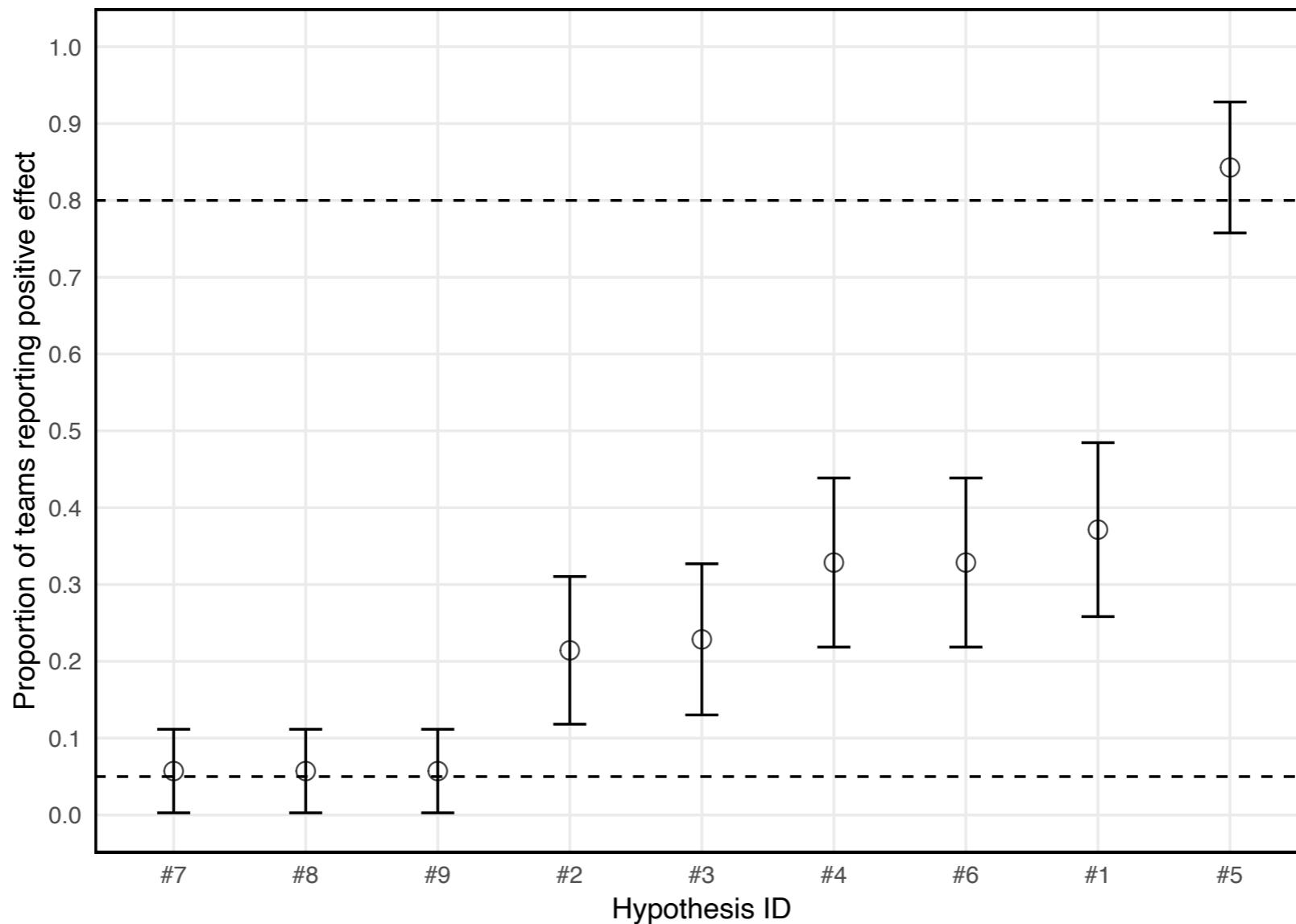
Best practices in data analysis and sharing
in neuroimaging using MRI

Thomas E Nichols¹, Samir Das^{2,3}, Simon B Eickhoff^{4,5}, Alan C Evans^{2,3}, Tristan Glatard^{2,6}, Michael Hanke^{7,8},
Nikolaus Kriegeskorte⁹, Michael P Milham^{10,11}, Russell A Poldrack¹², Jean-Baptiste Poline¹³, Erika Proal¹⁴,
Bertrand Thirion¹⁵, David C Van Essen¹⁶, Tonya White¹⁷ & B T Thomas Yeo¹⁸

NATURE NEUROSCIENCE VOLUME 20 | NUMBER 3 | MARCH 2017

- No 2 teams used an identical workflow
- Even with detailed written description it was often impossible to tell exactly what was done!

What is the effect of analytic variability on outcomes?

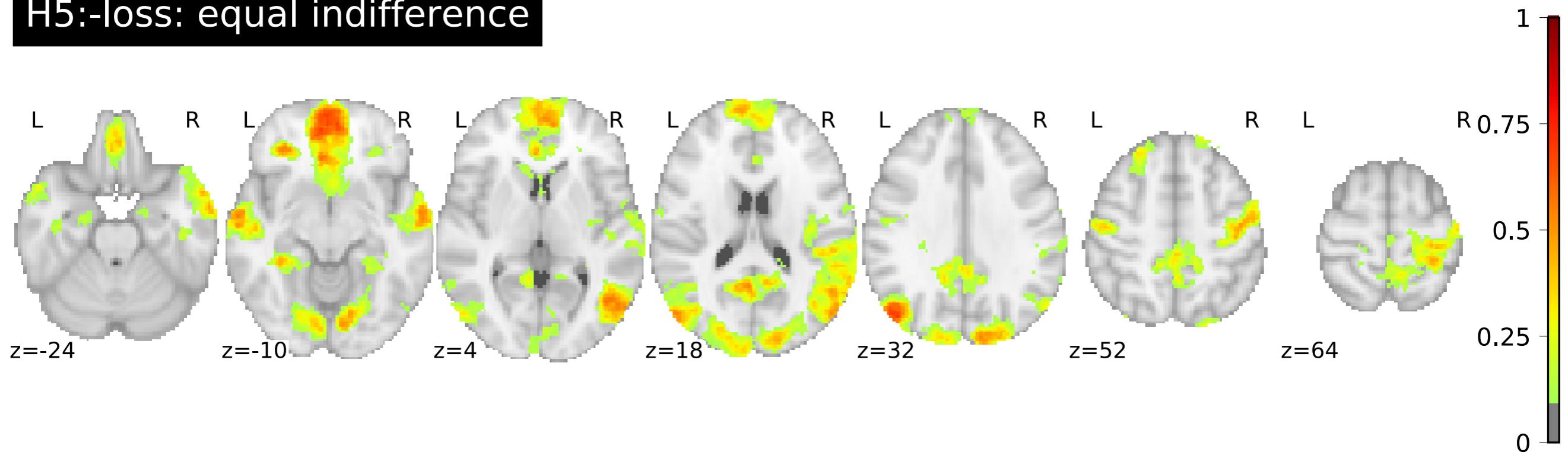


Across teams there were 33 different patterns of outcomes
For any hypothesis, there are at least 4 workflows that can
give a positive result

Variability of whole-brain results

Proportion of teams with activity in each voxel

H5:-loss: equal indifference

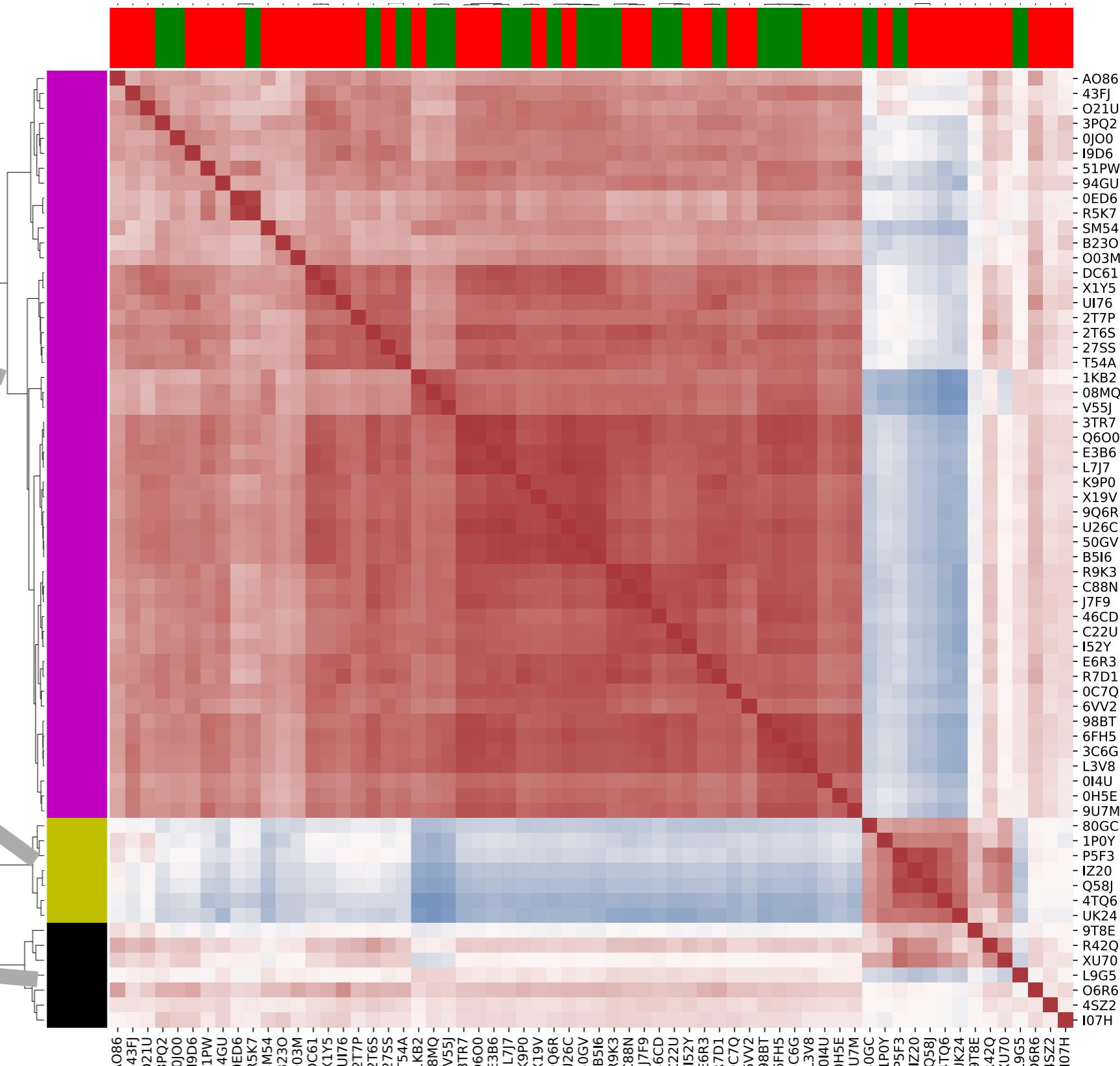
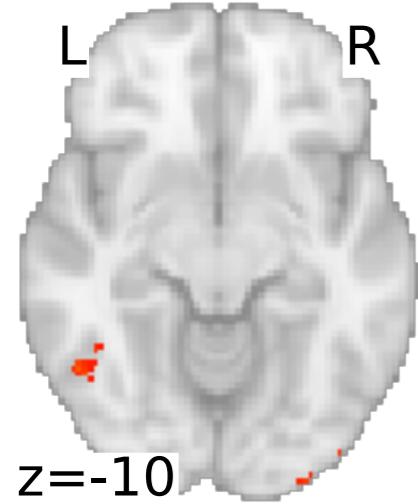
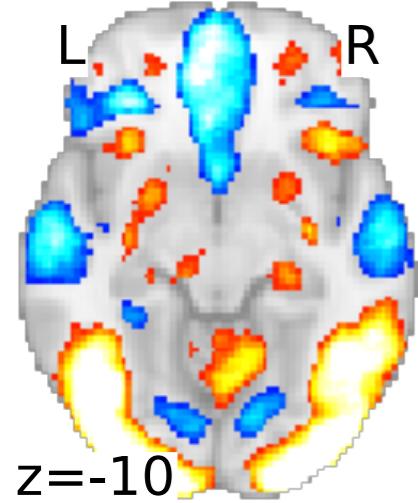
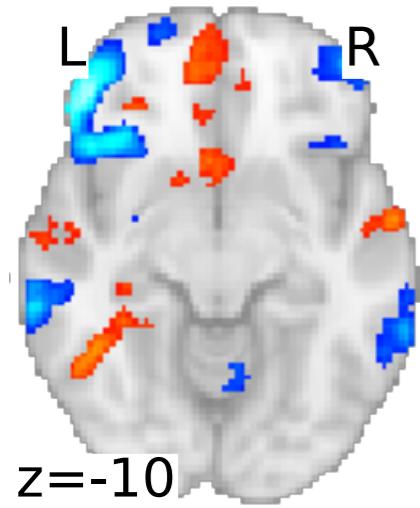


Maximum overlap for all hypotheses: 76%

Supplementary Table 5: Variability in the number of activated voxels reported across teams.

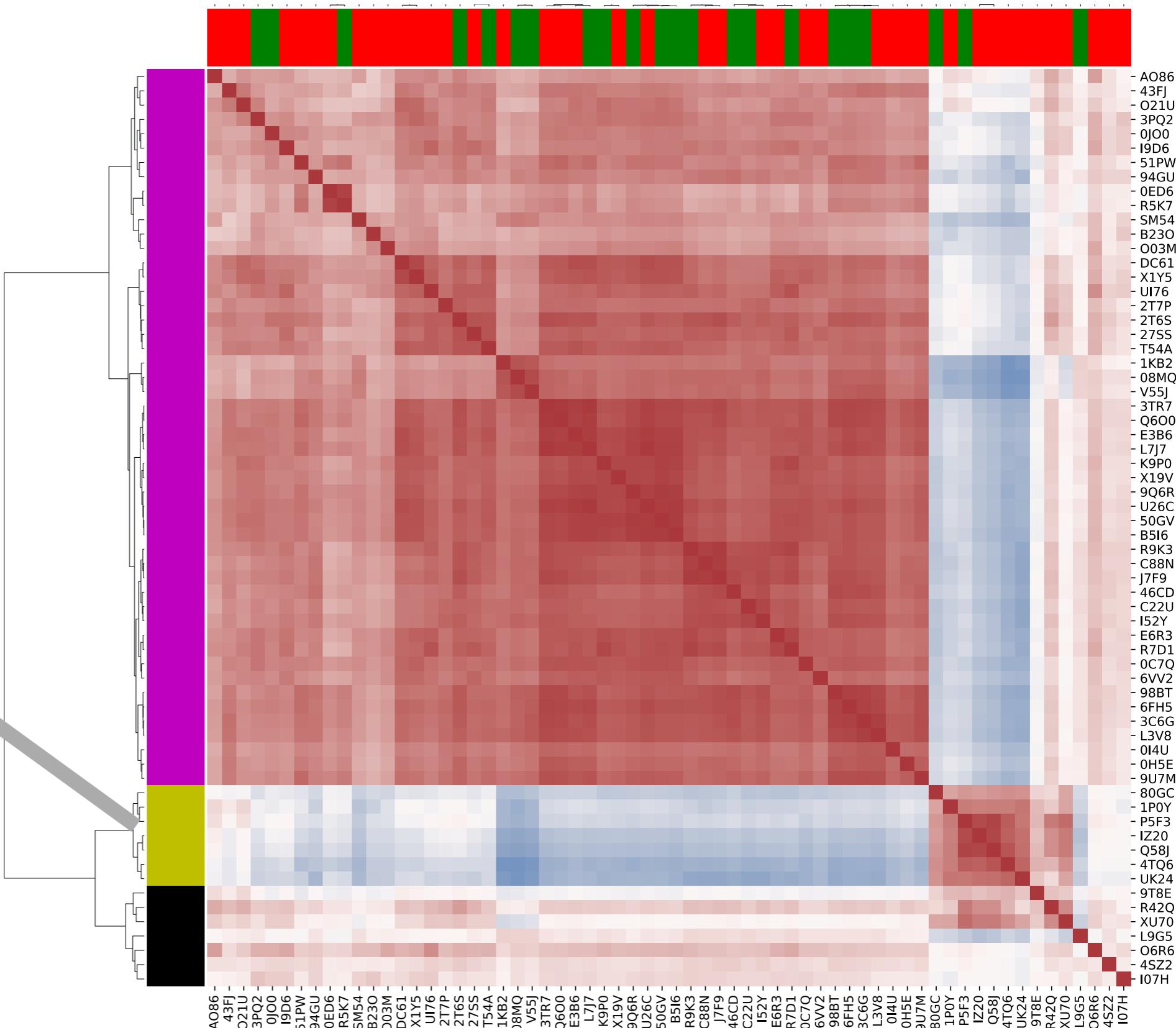
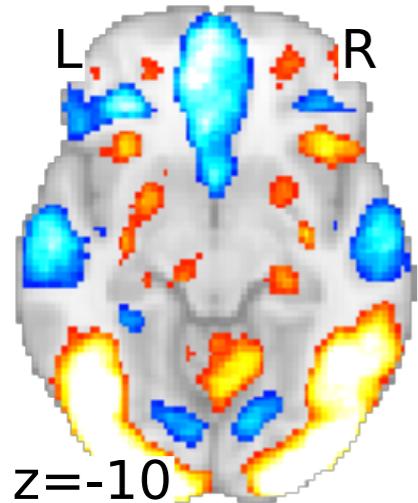
Hyp #	Minimum sig voxels	Maximum sig voxels	Median sig voxels	N empty images
1	0	118181	1940	8
2	0	135583	8120	2
3	0	118181	1940	8
4	0	135583	8120	3
5	0	76569	6527	11
6	0	72732	167	25
7	0	147087	9383	8
8	0	129979	475	16
9	0	49062	266	29

Hypothesis testing outcomes belie similarity of underlying statistics



Hypothesis testing outcomes belie similarity of underlying statistics

7 teams made
model mis-
specification
errors

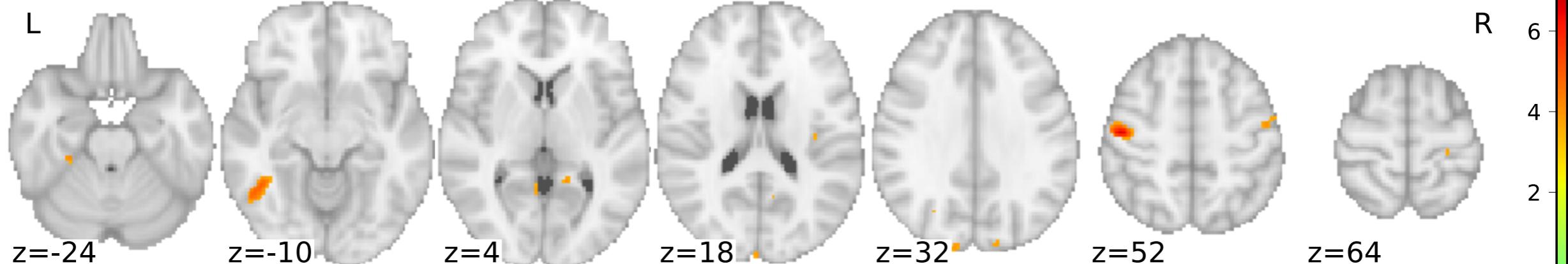


Meta-analysis across groups shows reliable findings

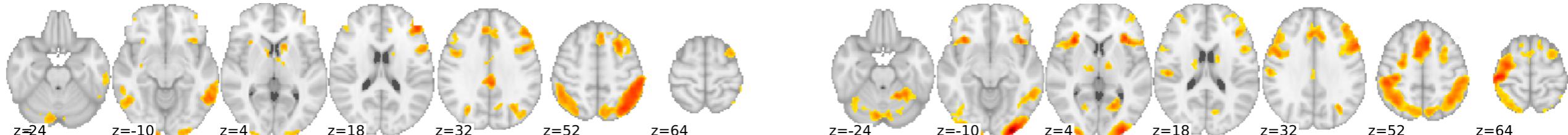
b

Image-based meta-analysis (IBMA) results

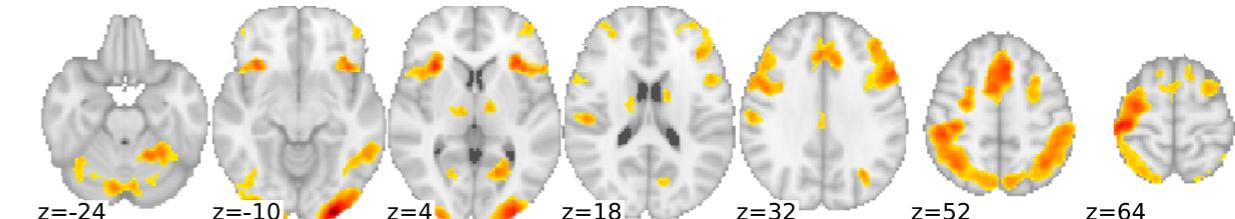
H1 + H3: +gain, equal indifference



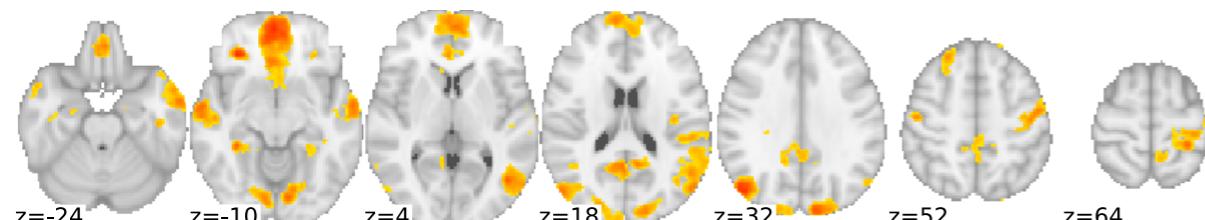
H2 + H4: +gain, equal range



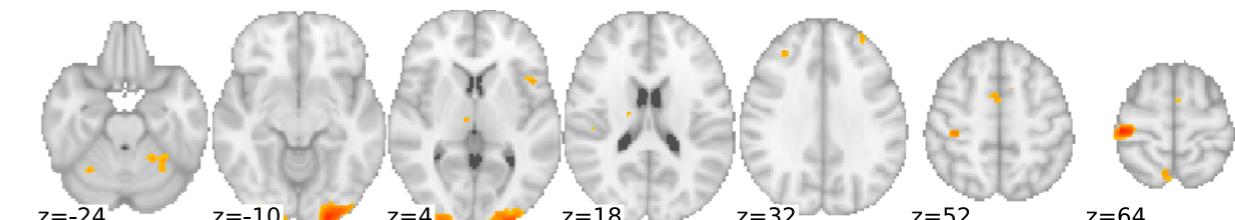
H7: +loss, equal indifference



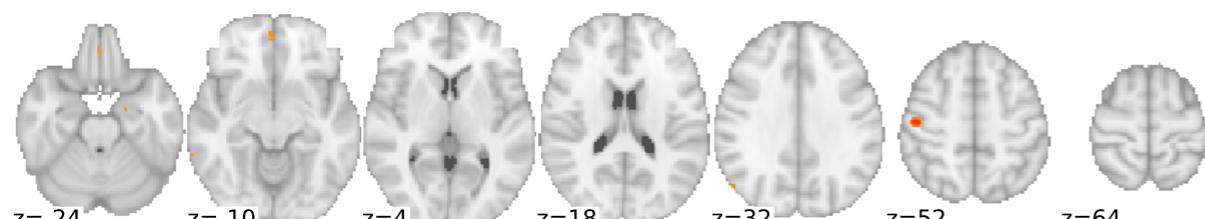
H5: -loss, equal indifference



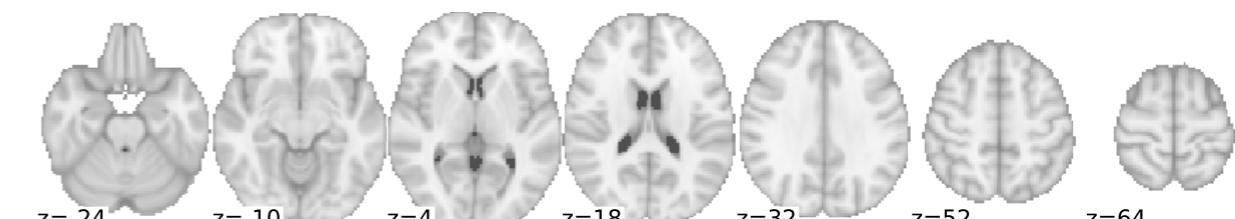
H8: +loss, equal range



H6: -loss, equal range



H9: +loss, ER > EI



Towards “multiverse” analysis

- Analytic variability has important impacts on scientific results
- Need to run data through multiple analysis workflows and assess convergence
- This will require automated model specification and execution using HPC
- Possible through Fitlins project using BIDS Statsmodels specification

BIDS Statsmodels

```
{  
  "Level": "Run",  
  "Transformations": [  
    {  
      "Name": "Factor",  
      "Inputs": ["trial_type"]  
    },  
    {  
      "Name": "Convolve",  
      "Model": "spm"  
      "Inputs": ["trial_type.A", "trial_type.B"]  
    },  
    {  
      "Name": "Lag",  
      "Inputs": ["rot_x", "rot_y", "rot_z",  
                "trans_x", "trans_y", "trans_z"],  
      "Outputs": ["d_rot_x", "d_rot_y", "d_rot_z",  
                  "d_trans_x", "d_trans_y", "d_trans_z"]  
    },  
    {  
      "Name": "Power",  
      "Order": 2,  
      "Inputs": ["rot_x", "rot_y", "rot_z",  
                 "trans_x", "trans_y", "trans_z",  
                 "d_rot_x", "d_rot_y", "d_rot_z",  
                 "d_trans_x", "d_trans_y", "d_trans_z"],  
      "Outputs": ["rot_x_2", "rot_y_2", "rot_z_2",  
                  "trans_x_2", "trans_y_2", "trans_z_2",  
                  "d_rot_x_2", "d_rot_y_2", "d_rot_z_2",  
                  "d_trans_x_2", "d_trans_y_2", "d_trans_z_2"]  
    },  
  ]  
}
```

- 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.

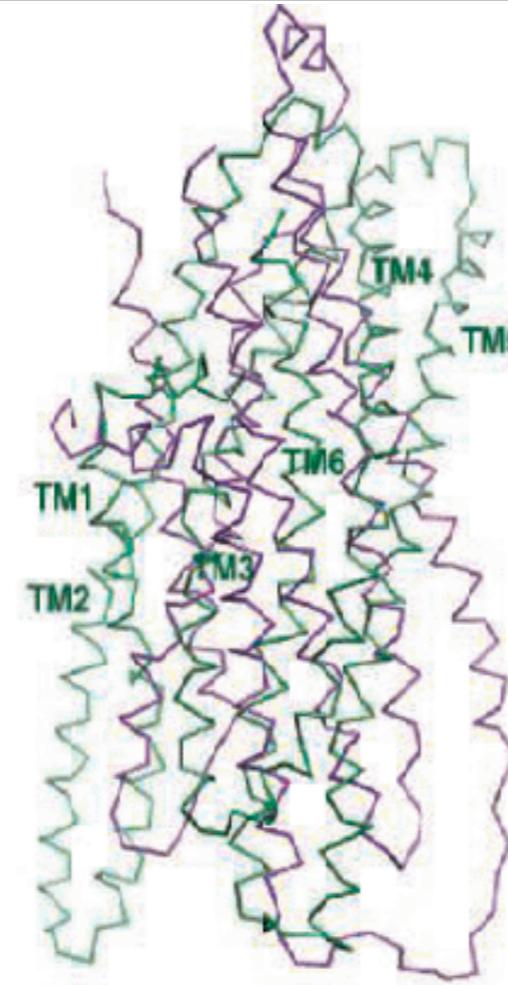
www.sciencemag.org SCIENCE VOL 293 7 SEPTEMBER 2001

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 ($Mg\text{-ADP-V}_i$) 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.

13 MAY 2005 VOL 308 SCIENCE www.sciencemag.org



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.

23 DECEMBER 2005 VOL 310 SCIENCE www.sciencemag.org

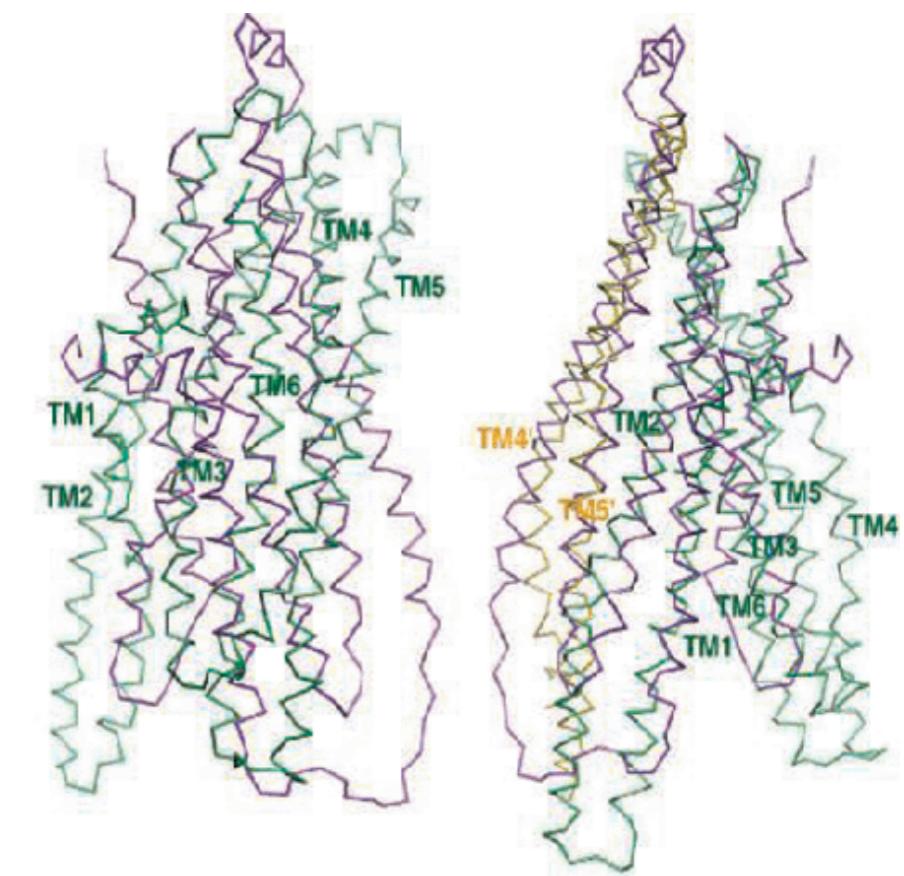
Threats to reproducibility: Software errors

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])           data[:,train]
        pred[test]=clf.predict(data[test])           data[:,test]
```

Results:
93% accuracy

Results:
53% accuracy

Software errors hit home

Cold
Spring
Harbor
Laboratory**bioRxiv**
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT |

Search

New Results

[Comment on this paper](#)

Design issues and solutions for stop-signal data from the Adolescent Brain Cognitive Development [ABCD] study

Patrick G. Bissett, McKenzie P. Hagen, Russell A. Poldrack

doi: <https://doi.org/10.1101/2020.05.08.204707>

This article is a preprint and has not been certified by peer review [what does this mean?].

However, for 1326 of the 7231 subjects (18.3% of all subjects), mean stop-failure RT was longer than mean RT on overt responses in go trials, suggesting that a significant subset of subjects violated context independence

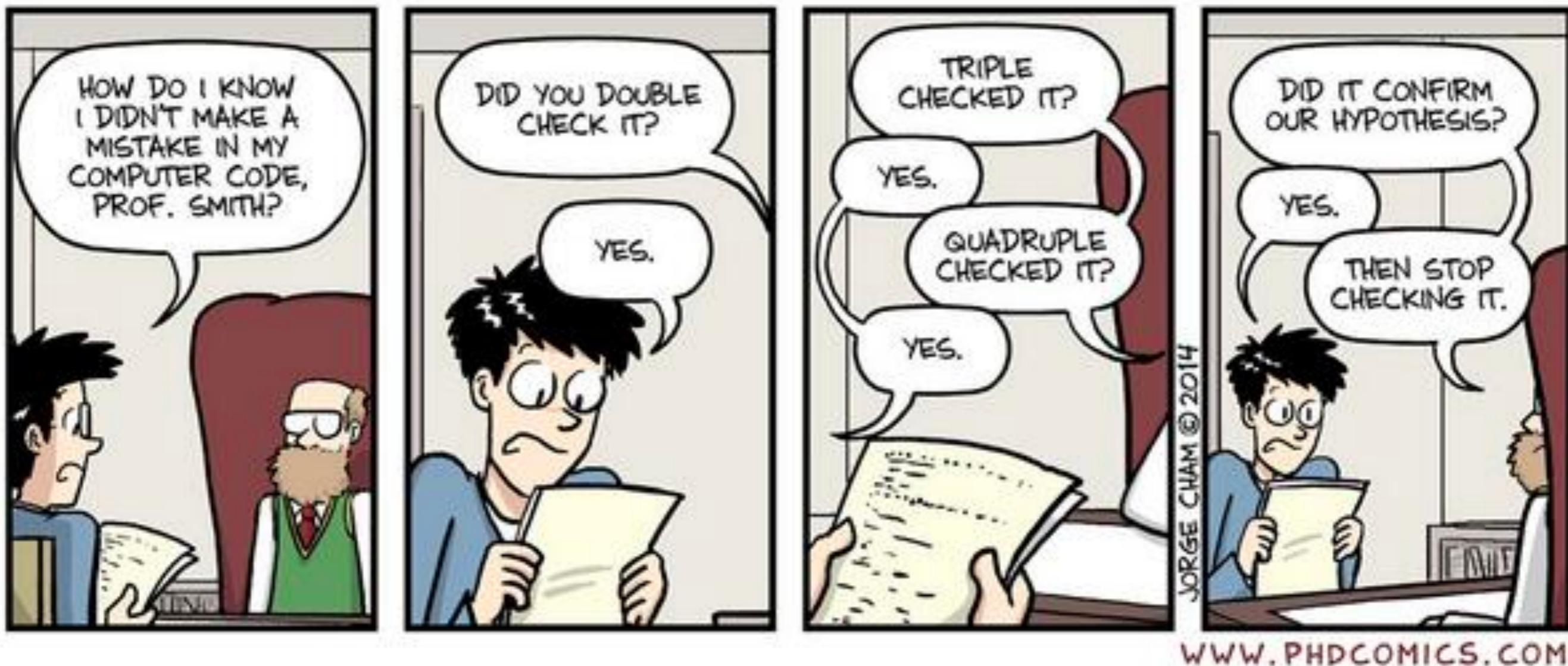
```
stop_resp_mask = ((~(SST_concat['StopSignal.RESP'].isnull()) &
                   ~(SST_concat['Fix.RESP'].isnull()) &
                   ~(SST_concat['SSD.RESP'].isnull())) &
                   (~SST_concat['stop_rt_adjusted'].isnull()))
```

Error identified by
Sage Hahn from ABCD
team



However, for 447 of the 7231 subjects (6.2% of all subjects), mean stop-failure RT was longer than mean RT on overt responses in go trials, suggesting that a subset of subjects violated context independence.

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”

Coding for reproducibility

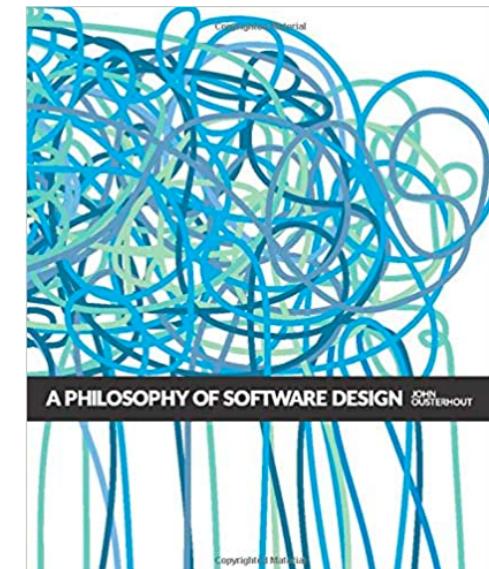
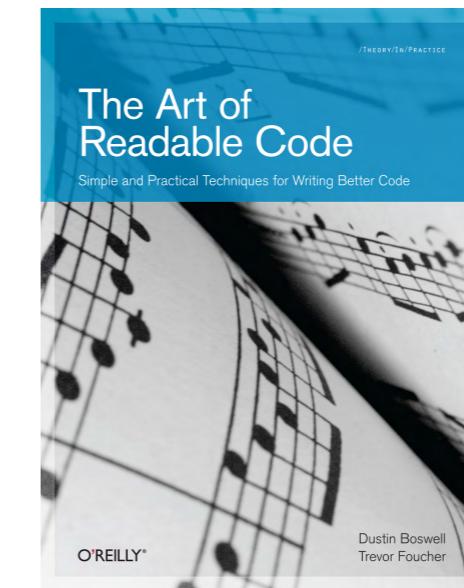
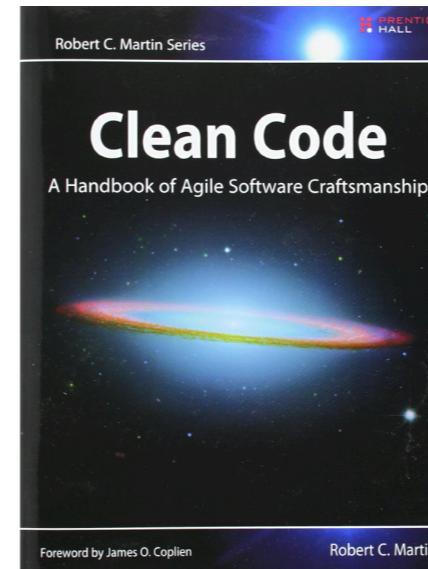
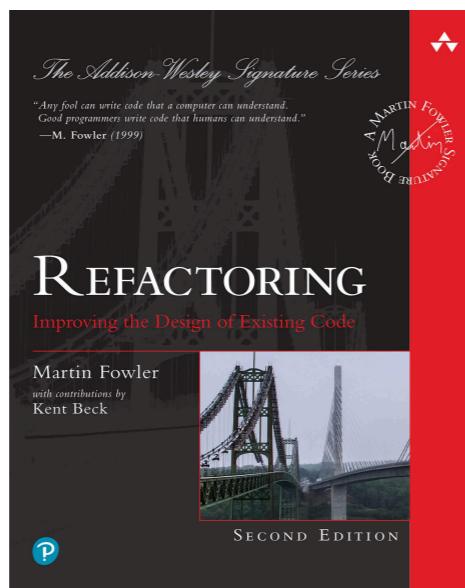
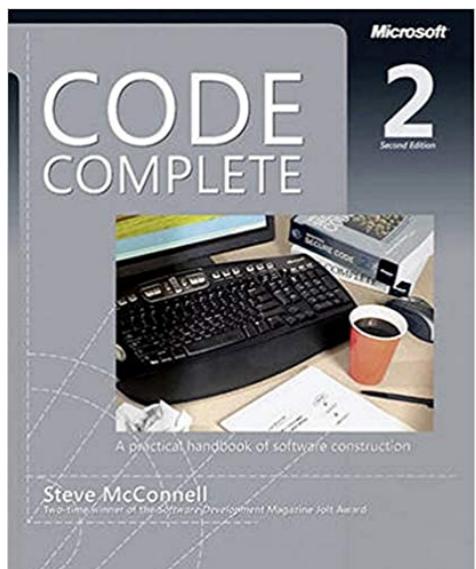
- High quality code should be:
 - Readable
 - Robust
 - Modular
 - Well-tested
- Coding is a craft
 - The only way to improve one's skills is through consistent and deliberate practice



<https://software-carpentry.org/>

Coding style

- “code is read much more often than it is written” so readability counts (G. van Rossum)
- Nearly all languages have conventions for coding style
 - Following them makes your code easier for others to understand
 - “others” includes your future self!
- There are also language-independent principles of software design



Commenting is not a magic balm for bad code

- Comments should help the reader to know as much as the writer knows about the code
 - Should express intention when it's not obvious
- “Clear and expressive code with few comments is far superior to cluttered and complex code with lots of comments” (Robert Martin, *Clean Code*)
- “Don’t comment bad code — rewrite it” (Kernighan & Plaugher)

Code review

- Linus's law
 - "given enough eyeballs, all bugs are shallow."

Code Review For and By Scientists

Marian Petre
Open University
m.petre@open.ac.uk

Greg Wilson
Mozilla Science Lab
greg@mozillafoundation.org

September 9, 2014

Based on our study, we believe the following should be priorities when scientists start doing code review:

1. Have a goal, a benefit in mind.
2. Start with a conversation: articulate your goals and expectations, build rapport.
3. Choose the right pieces of code for the first reviews. A good starting point is 3–4 pages long, fairly self-contained, and under active development, so that small patches are coming in regularly.
4. Make your code available in a repository with a typical data set and an overview of how the code works.
5. Set up a schedule and commit a little time on a regular basis.
6. Understand that it's the code that's being reviewed (not you). Tailor your comments to others on the same terms.

<https://arxiv.org/pdf/1407.5648.pdf>

Improvement: Software testing

- Smoke tests
 - Does everything run without exploding?
- Sanity checks
 - Are values within reasonable ranges?
- Unit tests
 - Does the function do what it is supposed to do?

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 (Github Actions, CircleCI, Travis)
- Automatically runs software tests whenever a new commit is pushed to Github

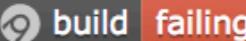
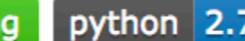
Improvement: Code quality analysis

- 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)
```

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

[build](#)  [build](#)  [codecov](#)  [circleci](#)  [python](#)  [python](#)  [pypi package](#)  [DOI](#) 

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>

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

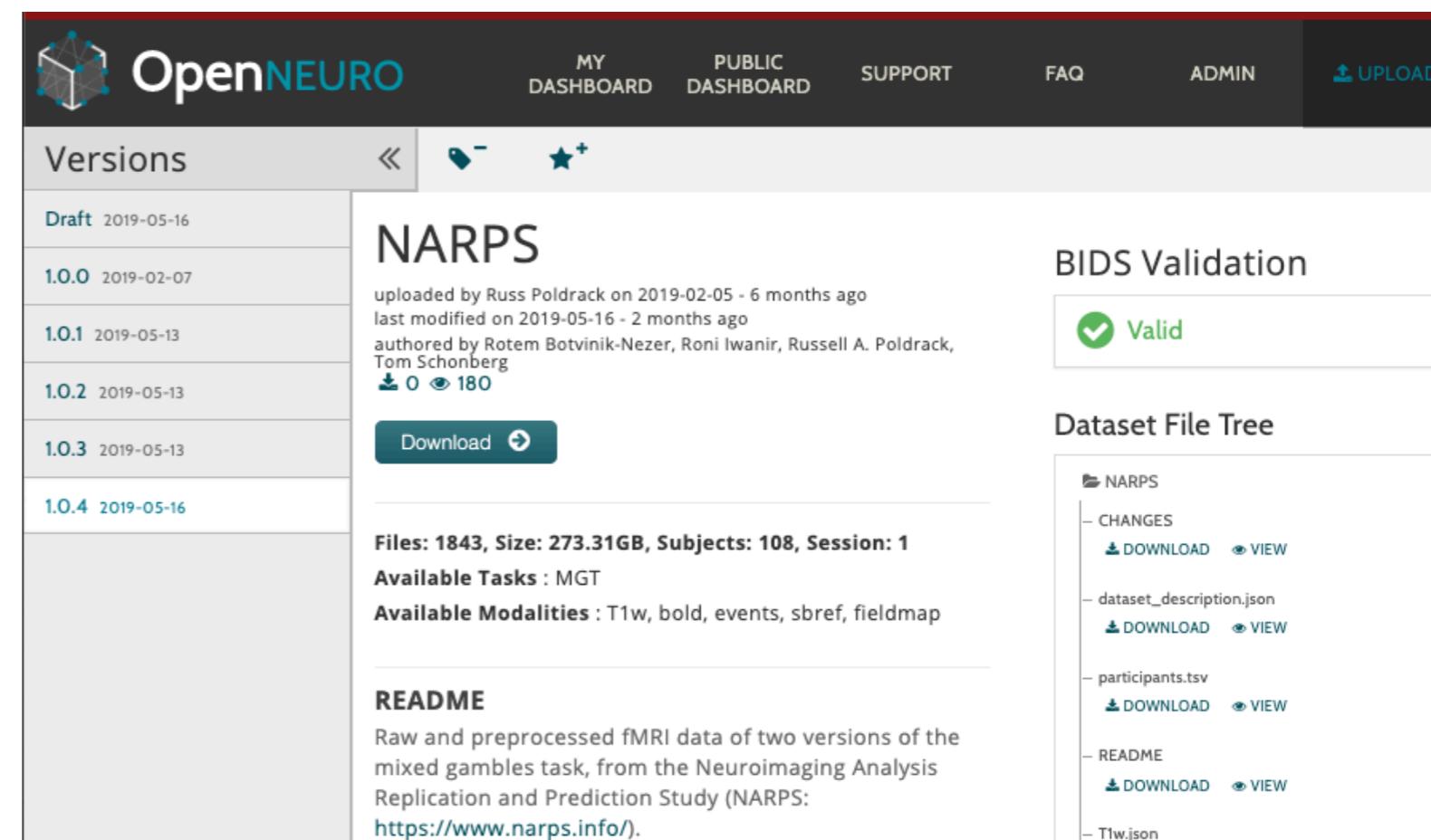
fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study

Received: 18 February 2019

Accepted: 7 June 2019

Published online: 01 July 2019

Rotem Botvinik-Nezer^{1,2}, Roni Iwanir^{1,2}, Felix Holzmeister^{ID 3}, Jürgen Huber^{ID 3}, Magnus Johannesson^{ID 4}, Michael Kirchler³, Anna Dreber^{ID 4,5}, Colin F. Camerer⁶, Russell A. Poldrack^{ID 7} & Tom Schonberg^{ID 1,2}



The screenshot shows the OpenNeuro dataset page for "NARPS". The top navigation bar includes links for "MY DASHBOARD", "PUBLIC DASHBOARD", "SUPPORT", "FAQ", "ADMIN", and "UPLOAD". The main content area has tabs for "Versions" (selected), "BIDS Validation" (Valid), and "Dataset File Tree".

Versions:

Version	Created
Draft	2019-05-16
1.0.0	2019-02-07
1.0.1	2019-05-13
1.0.2	2019-05-13
1.0.3	2019-05-13
1.0.4	2019-05-16

NARPS: uploaded by Russ Poldrack on 2019-02-05 - 6 months ago
last modified on 2019-05-16 - 2 months ago
authored by Rotem Botvinik-Nezer, Roni Iwanir, Russell A. Poldrack, Tom Schonberg
Download

BIDS Validation: Valid

Dataset File Tree:

- NARPS
 - CHANGES
 - dataset_description.json
 - participants.tsv
 - README
 - T1w.json

Files: 1843, **Size:** 273.31GB, **Subjects:** 108, **Session:** 1
Available Tasks: MGT
Available Modalities: T1w, bold, events, sbref, fieldmap

README: Raw and preprocessed fMRI data of two versions of the mixed gambles task, from the Neuroimaging Analysis Replication and Prediction Study (NARPS: <https://www.narps.info/>).

The importance of standards for data sharing

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

- » Data publication and archiving
- » Research data

Received: 18 December 2015

Accepted: 19 May 2016

Published: 21 June 2016

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

Krzysztof J. Gorgolewski¹, Tibor Auer², Vince D. Calhoun^{3,4}, R. Cameron Craddock^{5,6}, Samir Das⁷, Eugene P. Duff⁸, Guillaume Flandin⁹, Satrajit S. Ghosh^{10,11}, Tristan Glatard^{7,12}, Yaroslav O. Halchenko¹³, Daniel A. Handwerker¹⁴, Michael Hanke^{15,16}, David Keator¹⁷, Xiangrui Li¹⁸, Zachary Michael¹⁹, Camille Maumet²⁰, B. Nolan Nichols^{21,22}, Thomas E. Nichols^{20,23}, John Pellman⁶, Jean-Baptiste Poline²⁴, Ariel Rokem²⁵, Gunnar Schaefer^{1,26}, Vanessa Sochat²⁷, William Triplett¹, Jessica A. Turner^{3,28}, Gaël Varoquaux²⁹ & Russell A. Poldrack¹



- 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

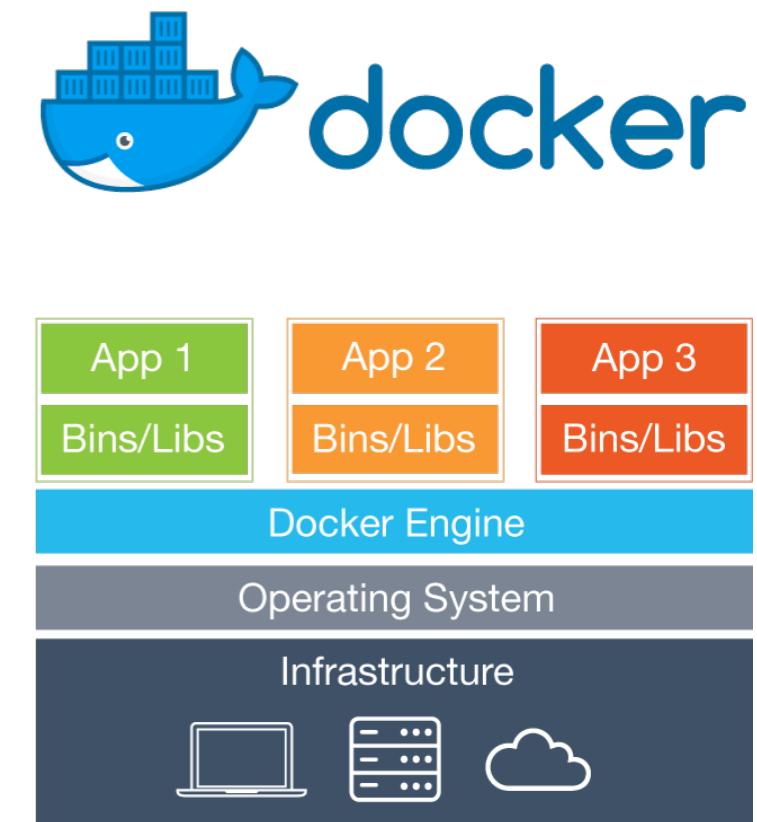
- dwi

- 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

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

The screenshot shows the CircleCI web interface for the 'poldrack' project. The left sidebar includes icons for Jobs, Workflows, Insights, Add Projects, Team, and Settings. The main area displays a table of jobs:

Project	Status	Description	Type	Time	Duration	SHA
psych10-book	SUCCESS	poldrack / narps / master #95 Merge pull request #26 from poldrack/develop	workflow	1 hr ago	25:58	c604014
Self_Regulation_Ontology	SUCCESS	poldrack / narps / develop #94 fix directory names	build	2 hr ago	19:15	a051ede
narps	FAILED	poldrack / narps / develop #93 add cache dirs	workflow	2 hr ago	00:32	f2d9044
reproducible-workflows	CANCELED	poldrack / narps / develop #92 fix dir name	build	2 hr ago	00:03	5301e3e
ezdiff	CANCELED	poldrack / narps / develop #91 fxx file name	workflow	2 hr ago	00:00	67faa1e
python-circleci						
r_testing						
BBSRC-git-demo						

CI now easily integrated with Github using Github Actions

Automated code quality assessment (codacy.com)

- Automatically checks code upon Github push

Issues breakdown

narps master

Badges Share

A Project certification

Quality evolution

Last 7 days Last 31 days

Issues 8% 26% Complex Files 0% = Duplicated code 25% ~15% Coverage -

-- Trend for the next 31 days -- Pull request prediction - Quality standard

Days: 22, 23, 24, 25, 26, 27, 28, 29, 30, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, JULY

27 total issues

Category	Total
Security	17
Error Prone	1
Code Style	5
Compatibility	0
Unused Code	4
Performance	0

See all Issues

Current Issues master

Filter All languages All categories All levels All paths

ImageAnalyses/CheckSimulatedValues.py

Unused NarpsDirs imported from narps

```
9 from narps import Narps, NarpsDirs, hypnums # noqa, fl
```

Use of assert detected. The enclosed code will be removed when co

```
16 assert os.path.exists(simdir)
```

Validation testing for data analysis code

- Parameter recovery
 - generate synthetic data with a known answer, and make sure that your code can recover the known answer
 - Best to do this before ever touching the real data!
- Null/randomization testing
 - Generate data with no effect (either synthetic or via randomization of real data) and make sure that no effect is found
- There are increasingly powerful frameworks for generating simulated data
 - Python: Synthetic Data Vault (<https://sdv.dev/SDV/>)

Validation for NARPS study

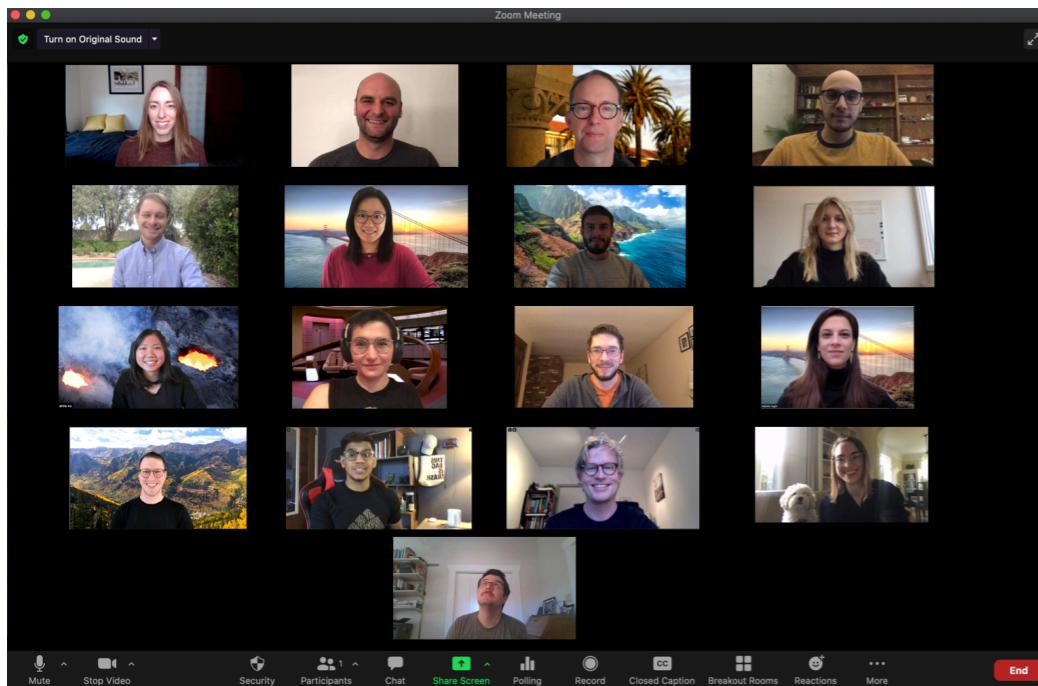
- Generate simulated data for each analysis team
 - Generate simulated data based on known maps
 - Include some teams with either completely random data or anticorrelated signal
 - Include different levels of variance across teams
- Run entire analysis stream on the simulated data
- Compare results of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

SUCESSS: all hypotheses correlated > 0.95

Conclusion

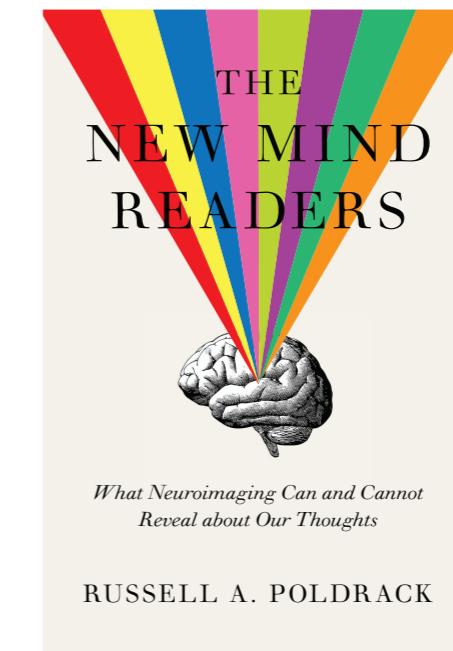
- Neuroimaging has a reproducibility problem
 - Likely not very different from other areas of science with complex computational workflows and large datasets
- Statistical power is key to solving the problem
- Software engineering tools can help ensure reproducibility of complex analyses
- The tools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

Acknowledgments



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

Many awesome collaborators
(too many to name!)



<http://statsthinking21.org>