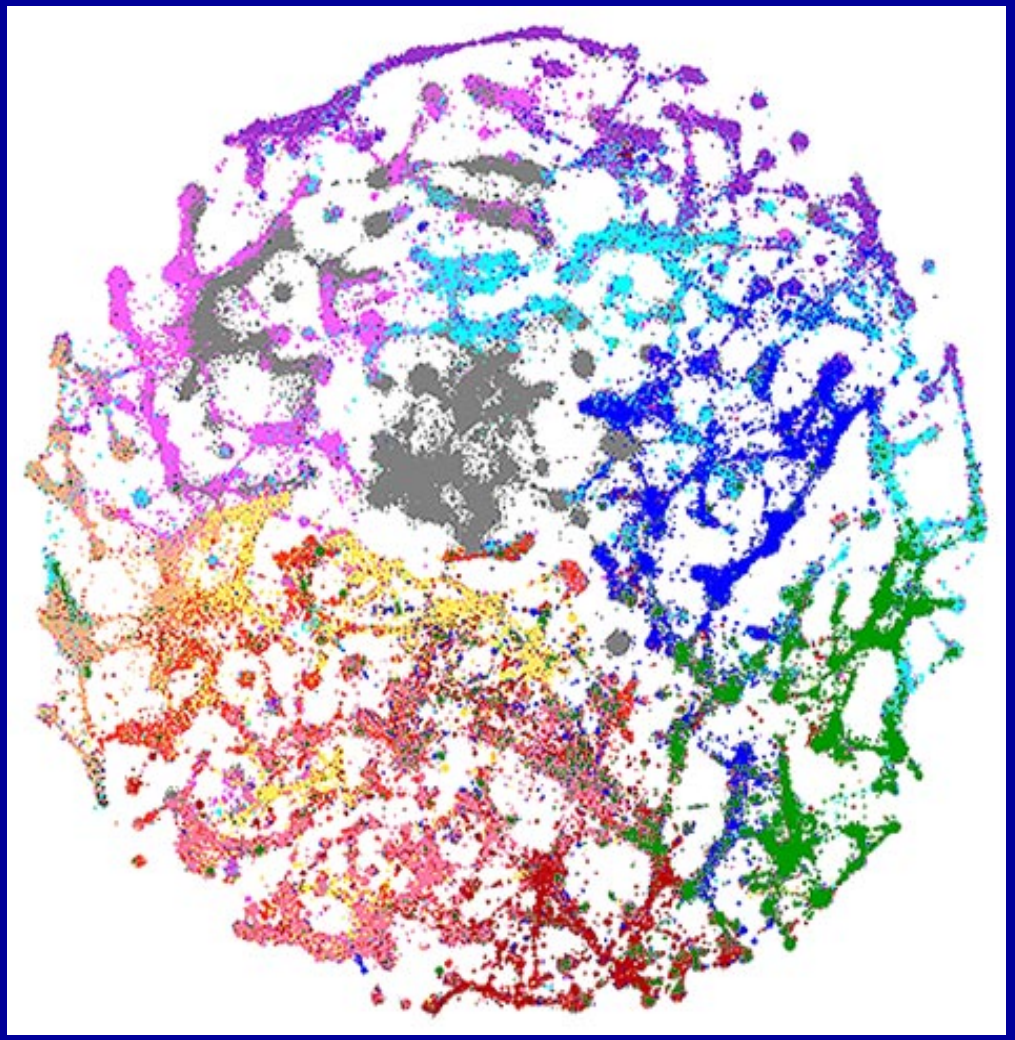
What Have We (Not) Learnt from Millions of Scientific Papers with p-values?

John P.A. Ioannidis, MD, DSc

C.F. Rehnborg Chair in Disease Prevention

Professor of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University

Co-Director, Meta-Research Innovation Center at Stanford (METRICS)



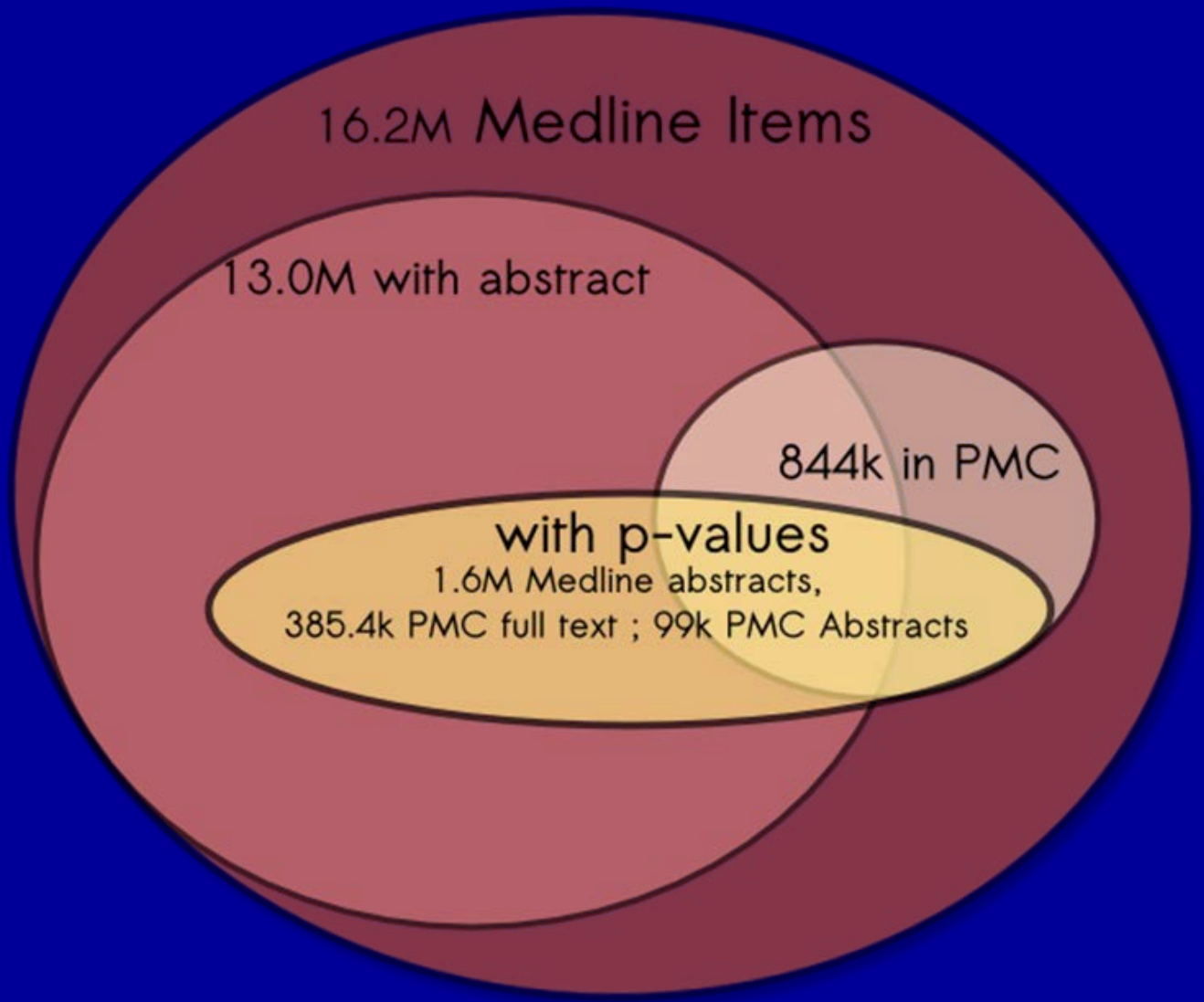
Different scientific disciplines use different statistical inference tools by tradition (not necessarily justified)

A map of recent science: 20 million papers, 2 million patents, 200000 clusters lasting 2-16 years each

NHST and p-value thresholds

- 0.05 – highly prevalent in biomedical and social sciences
- $3x10^{-7}$ (5 sigma) – in high energy physics
- $5x10^{-8}$ (genome-wide significance) – in genome epidemiology

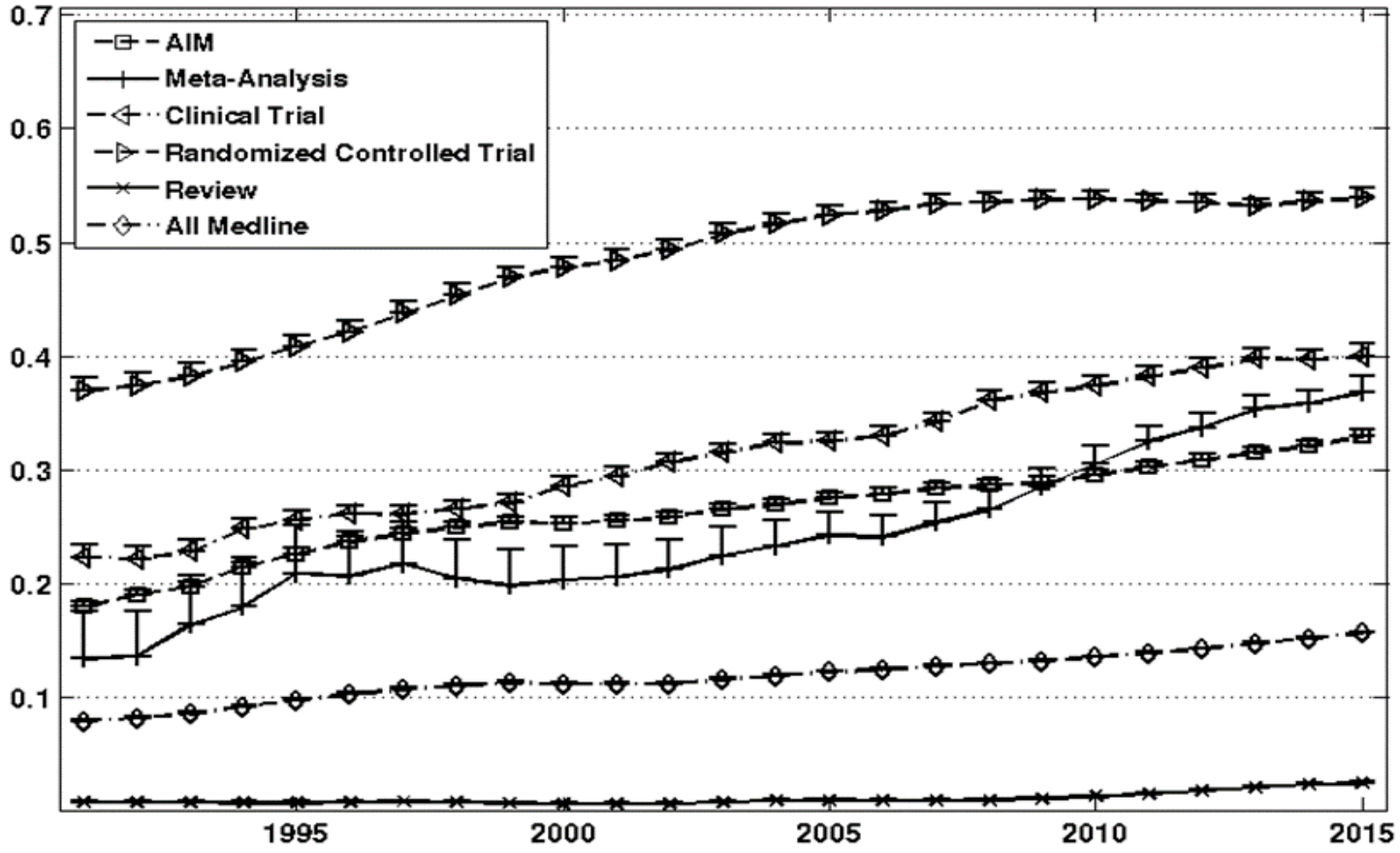
A view of the biomedical literature



Proportion of PMC papers with P-values in their abstract or text

- Core clinical journals: 78.4%
- Meta-analyses: 82.8%
- RCTs: 76.0%
- Clinical trials (excluding RCTs): 75.7%
- Reviews (excluding meta-analyses): 22.3%
- All papers: 51.1%

The proportion of PubMed items that have any P-values in the abstract is increasing

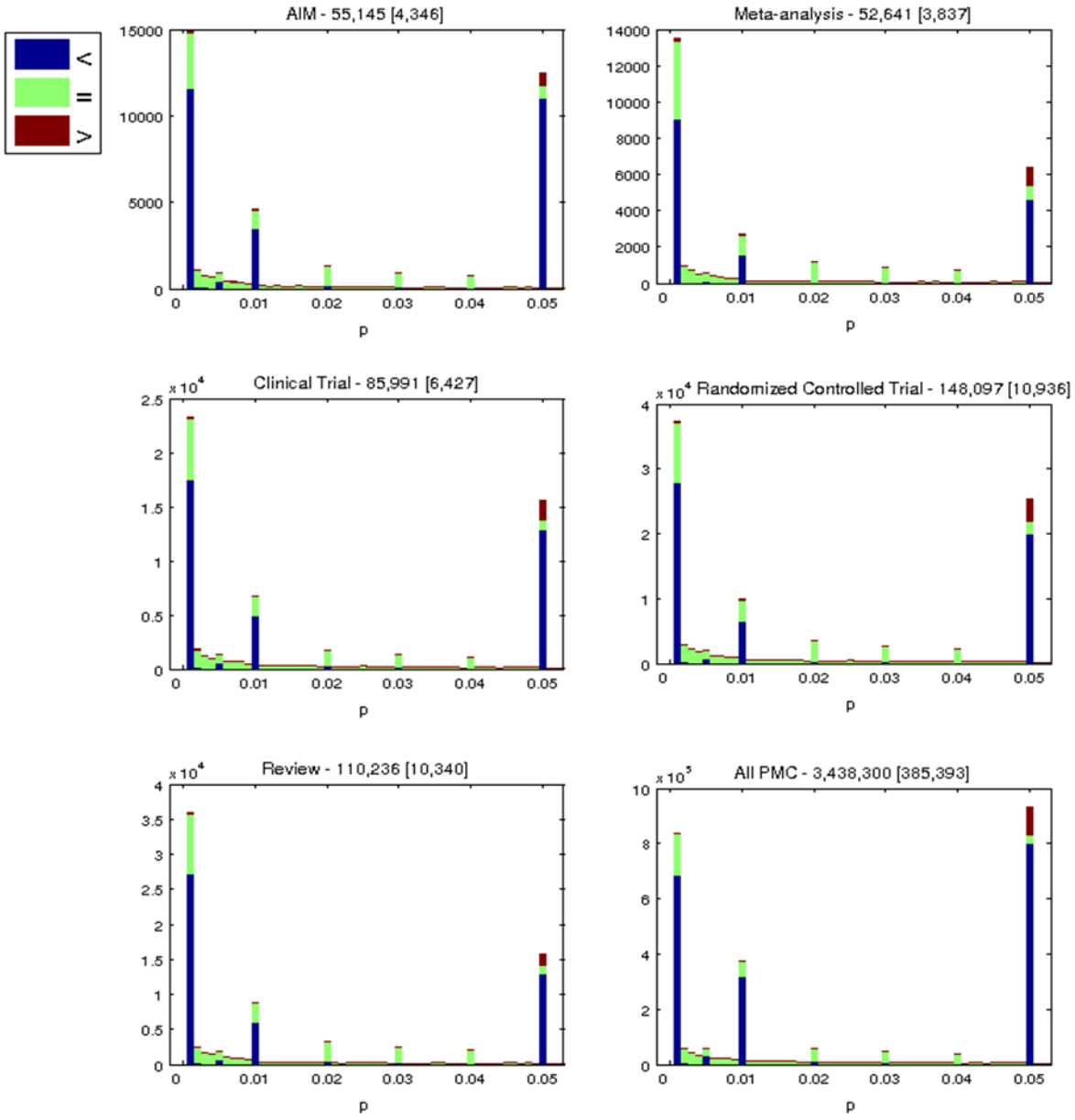


10^x (very small) p-values are overall uncommon

Format	Number of occurrences in 1990-2015	Average -log(P-value)
Plain	4,663,091	2.03
10^x	21,708	8.9
Percentage	1,042	1.78
Typo?	3,156	0.82
Exp^x	1,843	9.27

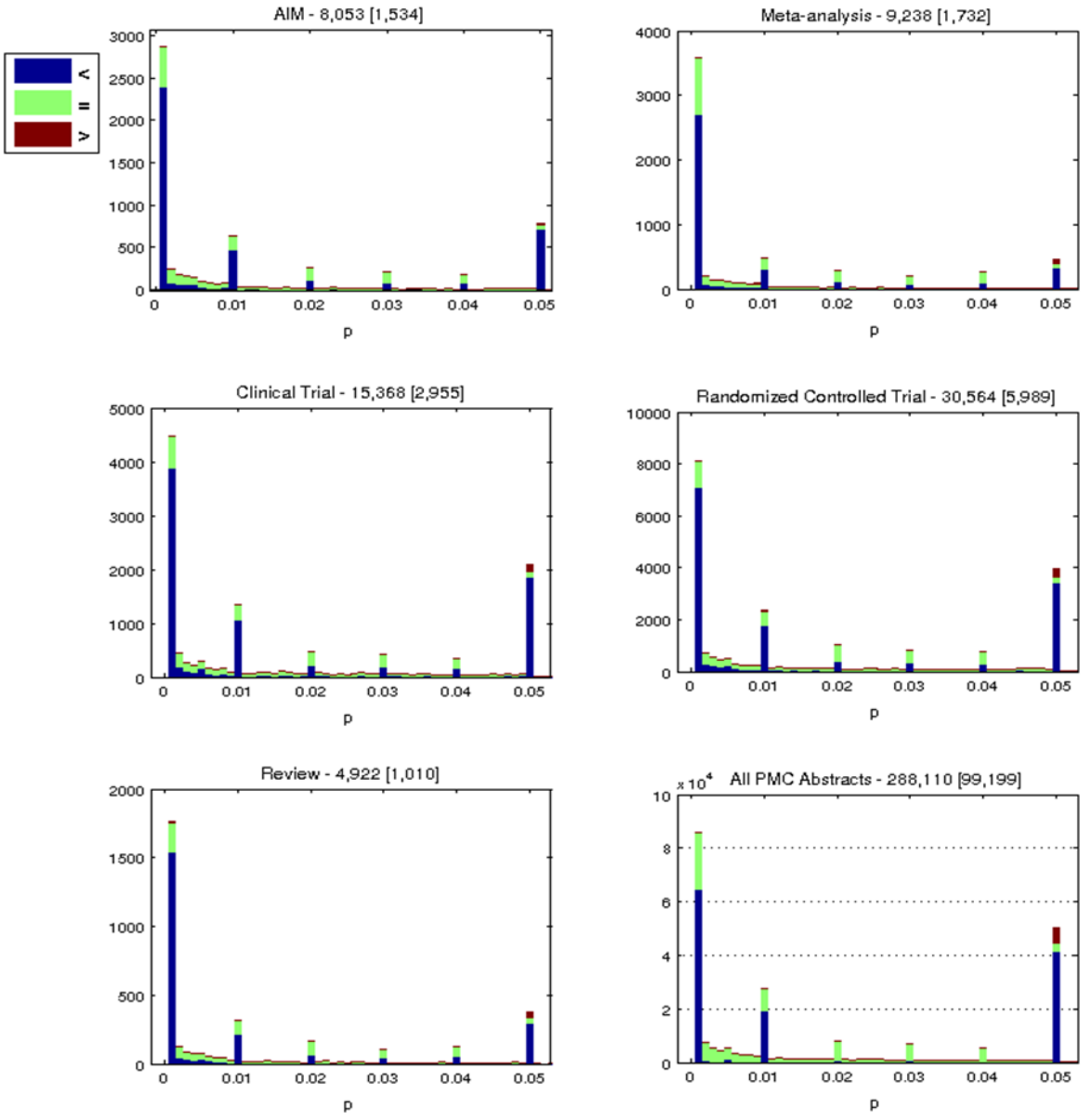
3,428,300 P-values in 385,393 PMC full-text papers that have abstracts

Across the entire literature, there are more p-values at or near 0.05 than in any other bin

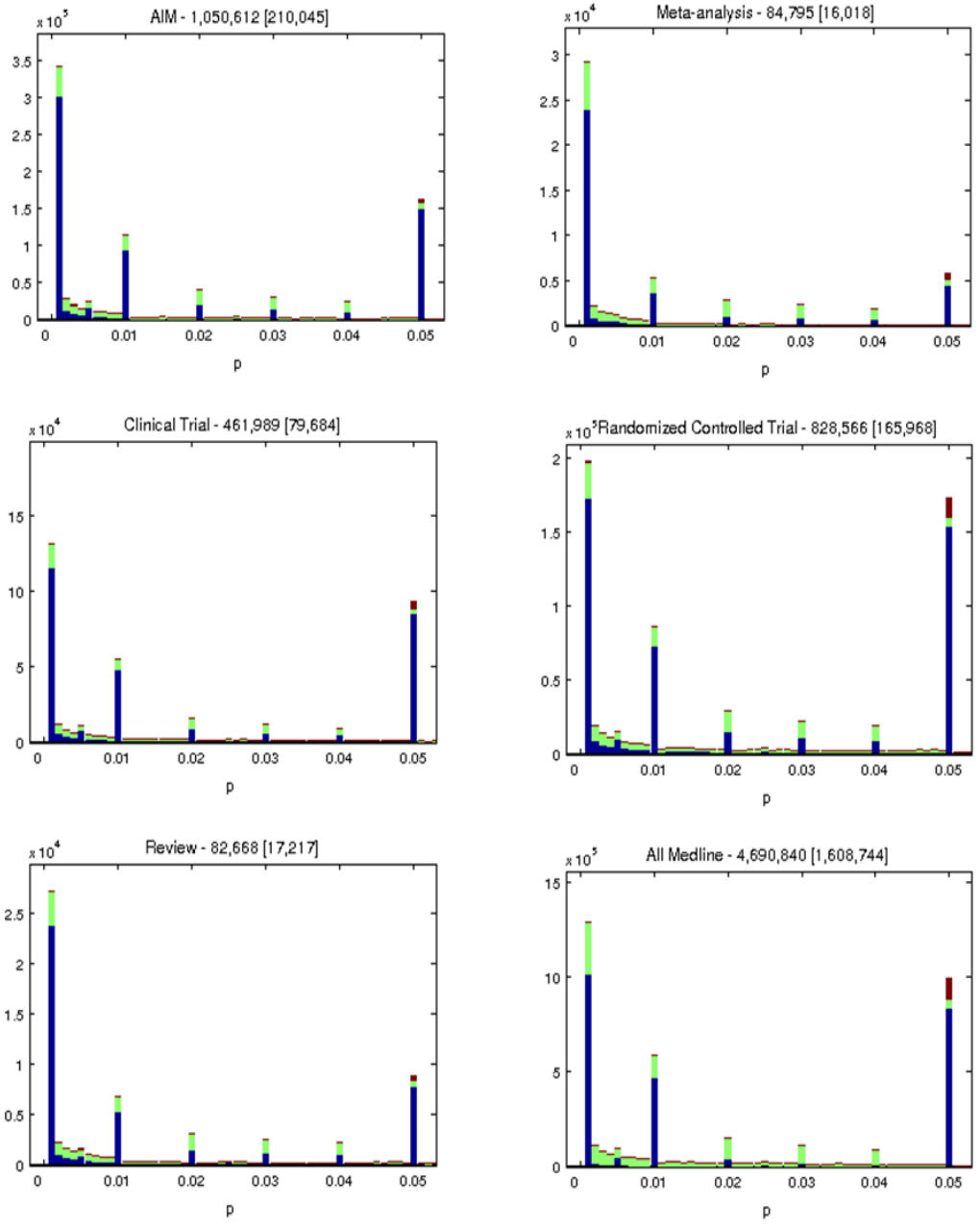


288,110 P-values in the abstracts of the 88,307 PMC papers that have P-values (among the 385,393)

Abstracts select more prominent P-values to report



4,690,840 P-values in the abstracts of 1,608,744 Medline papers with P-values



R=number of 0.05 P-values/number of 0.001 or less P-values

- For the P-values in the full-texts of PMC articles, R=1.11
- For the P-values in the abstracts of the respective papers, R=0.59

More impressive P-values are far more prevalent in the abstracts than in the full-texts

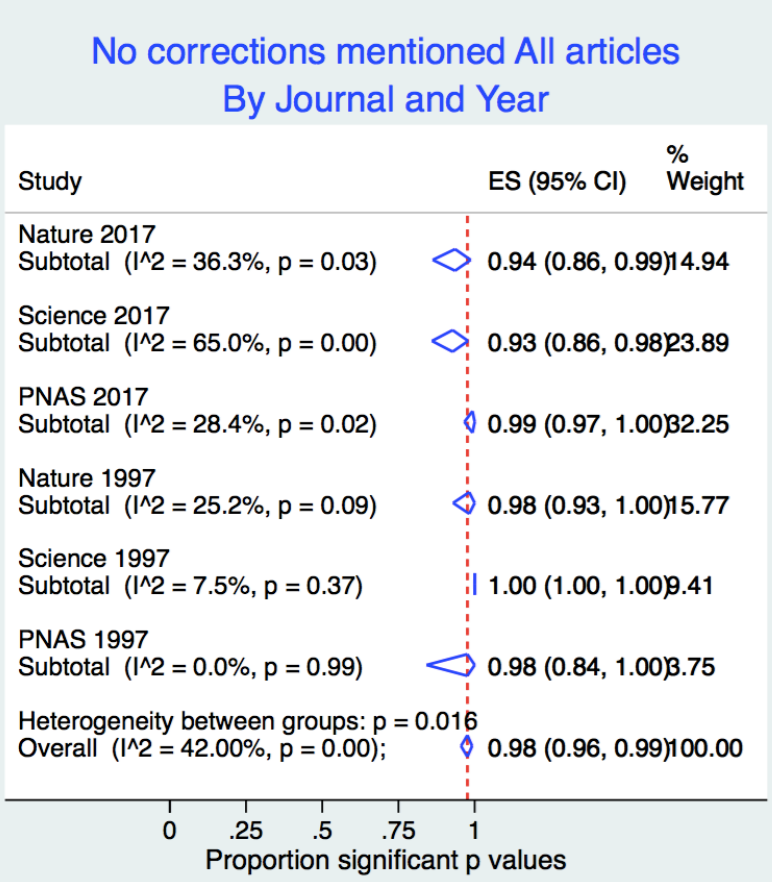
Comparison of R in abstracts and full-texts

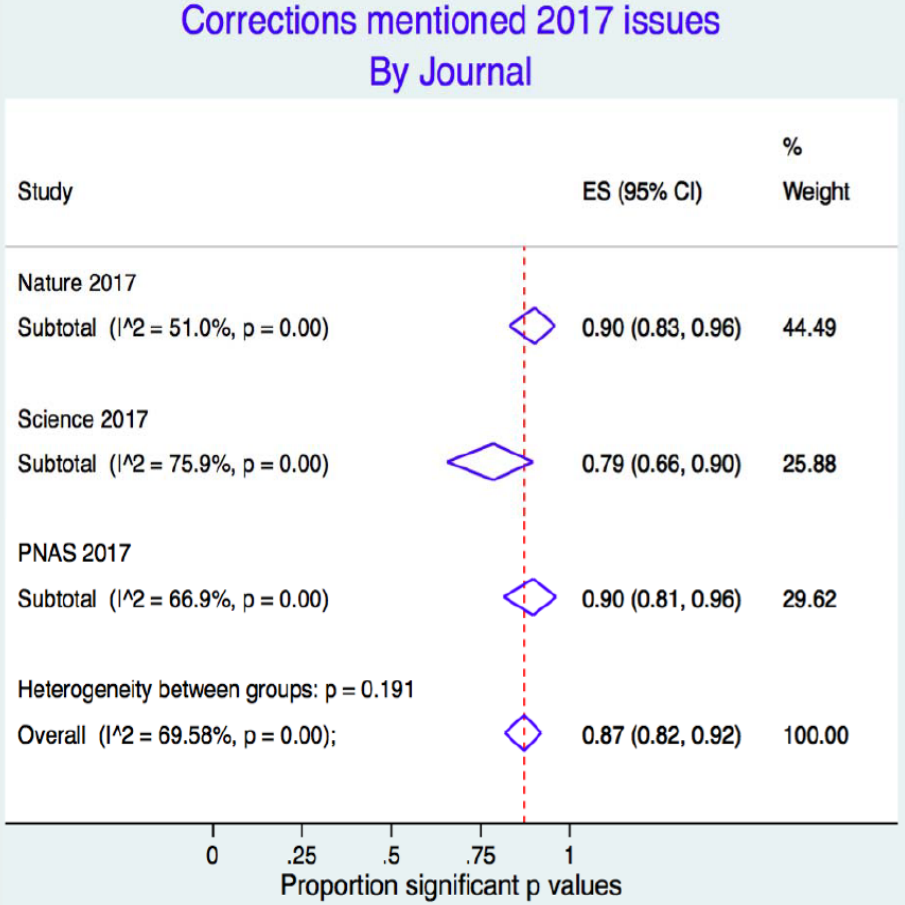
Full-text	Abstracts
Meta-analyses R=0.47	R=0.13
Reviews R=0.44	R=0.21
RCTs R=0.68	R=0.49
Clinical trials R=0.67	R=0.47
Core clinical journals R=0.84	R=0.27
All papers R=1.11	R=0.59

Layers of selection

- All p-values obtained in analyses
- P-values selected for presentation in tables and figures
- P-values discussed in the text
- P-values selected for presentation in the abstract
- P-values used for making inferences/conclusions

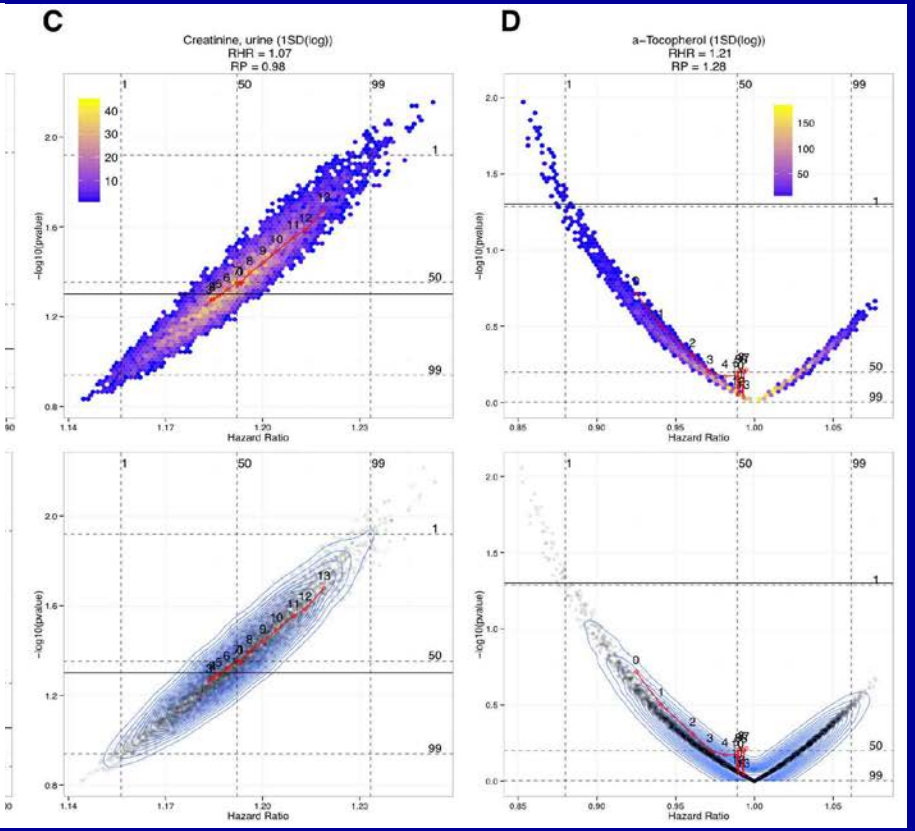
P-values in tables and figures of Nature, Science, and PNAS



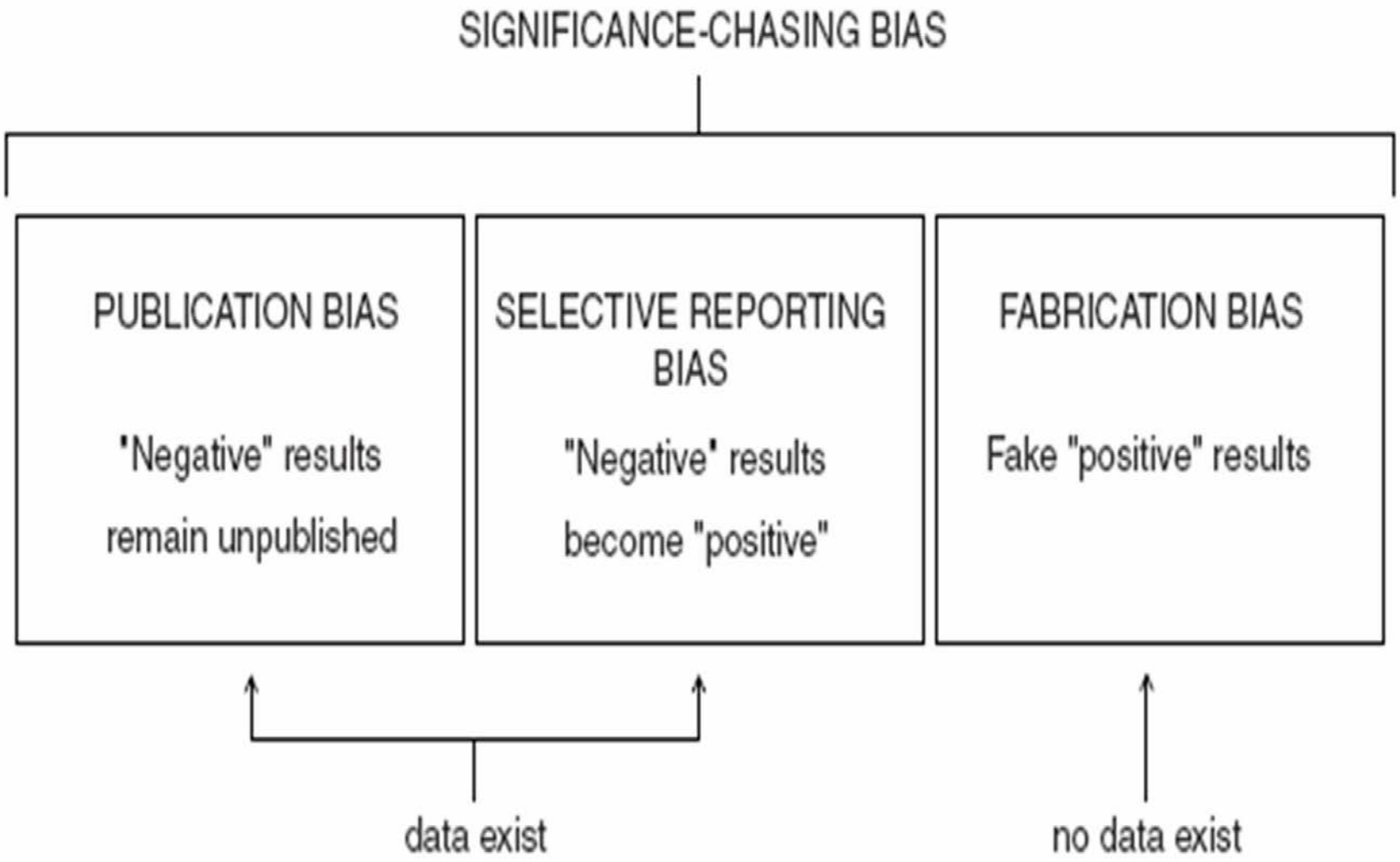


Cristea, Ioannidis, in preparation

Almost any result can be obtained: Vibration of effects and the Janus phenomenon



Patel, Burford, Ioannidis. JCE 2015; Patel and Ioannidis, JAMA 2015



Ioannidis PLoS Clinical Trials 2006 and Clinical Trials 2007

Journal of Experimental Psychology: General
2014, Vol. 143, No. 2, 534–547

0096-3445/14/$12.00 DOI: 10.1037/a0033242

P-Curve: A Key to the File-Drawer

Uri Simonsohn
University of Pennsylvania

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

Because scientists tend to report only studies (publication bias) or analyses (*p*-hacking) that "work," readers must ask, "Are these effects true, or do they merely reflect selective reporting?" We introduce *p*-curve as a way to answer this question. *P*-curve is the distribution of statistically significant *p* values for a set of studies (*p*s $<$ .05). Because only true effects are expected to generate right-skewed *p*-curves—containing more low (.01s) than high (.04s) significant *p* values—only right-skewed *p*-curves are diagnostic of evidential value. By telling us whether we can rule out selective reporting as the sole explanation for a set of findings, *p*-curve offers a solution to the age-old inferential problems caused by file-drawers of failed studies and analyses.

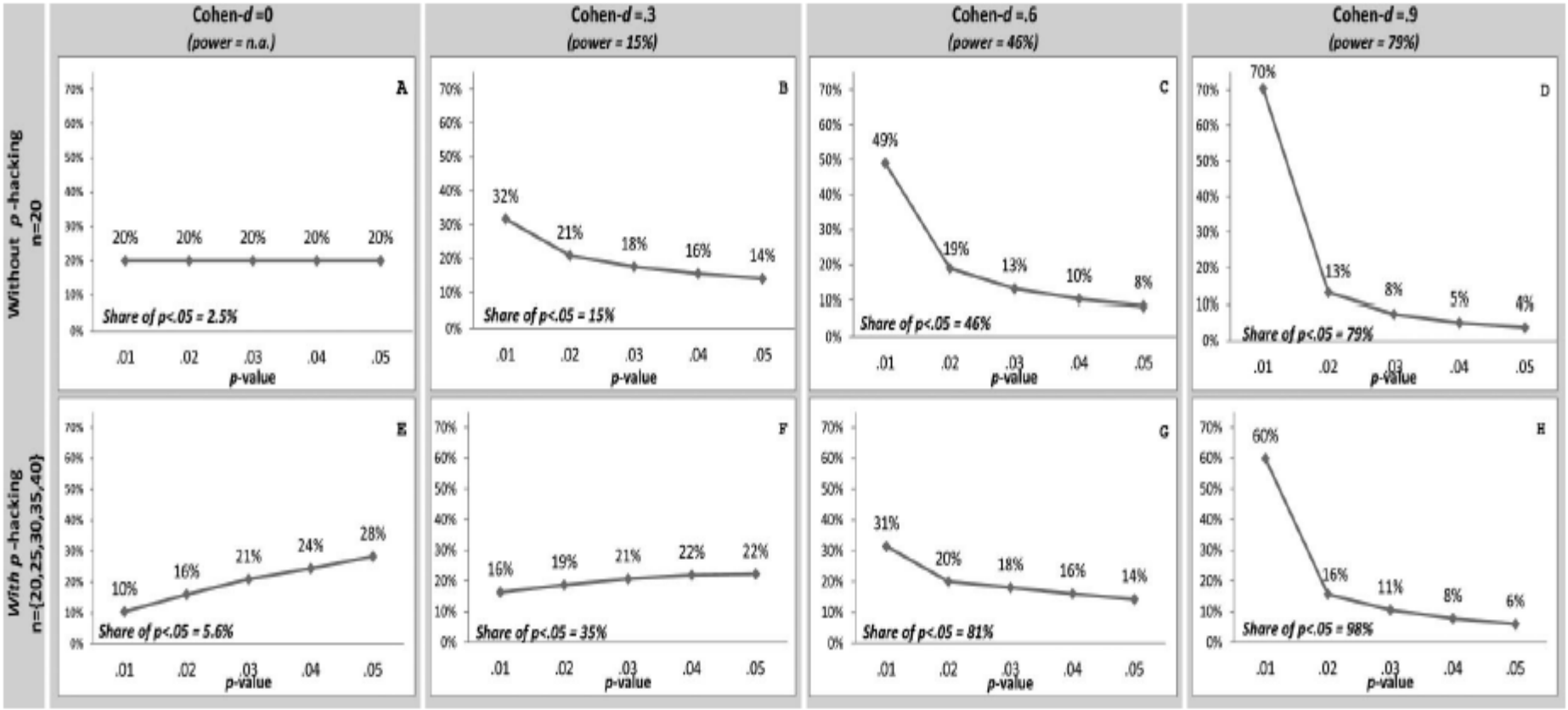
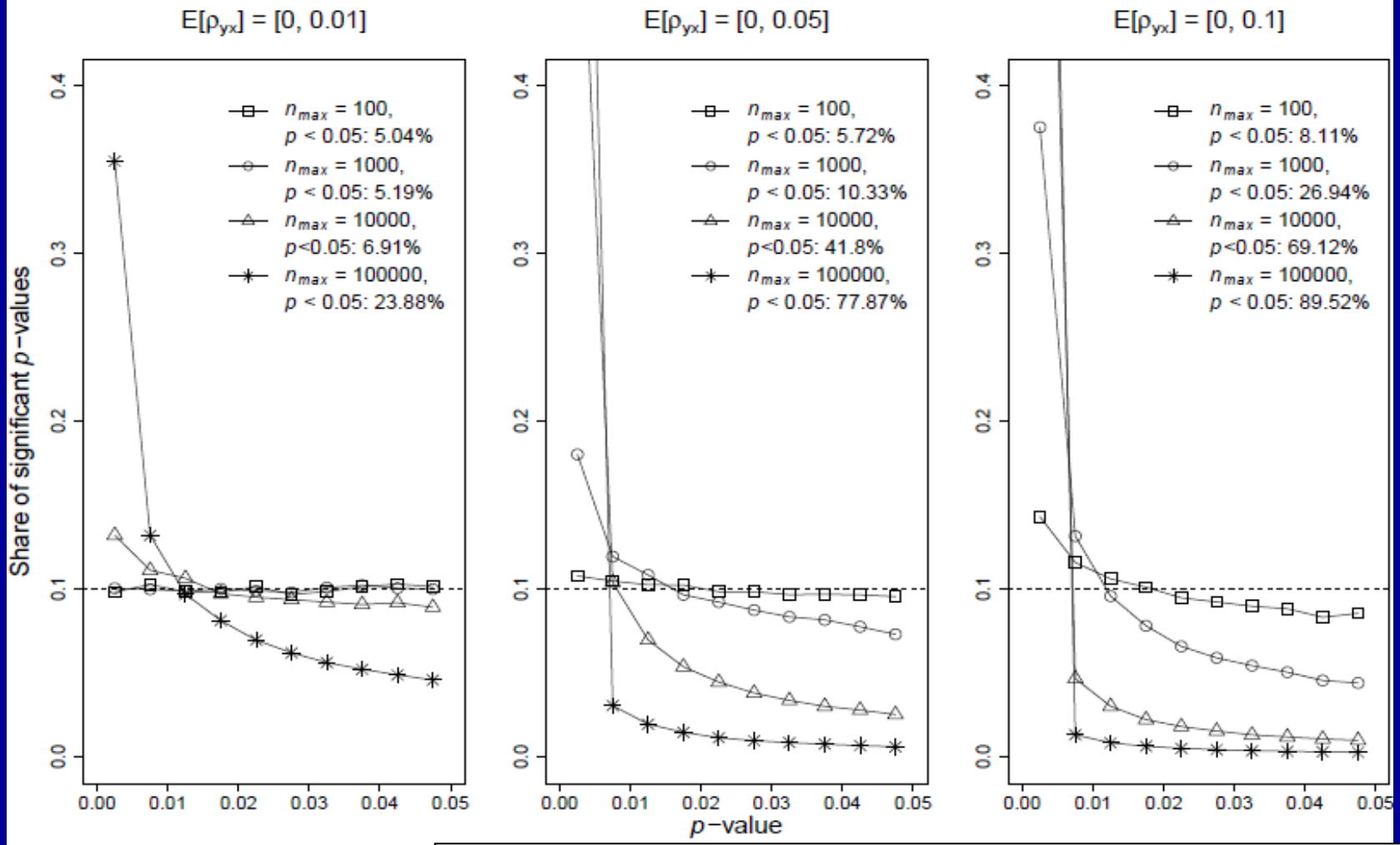


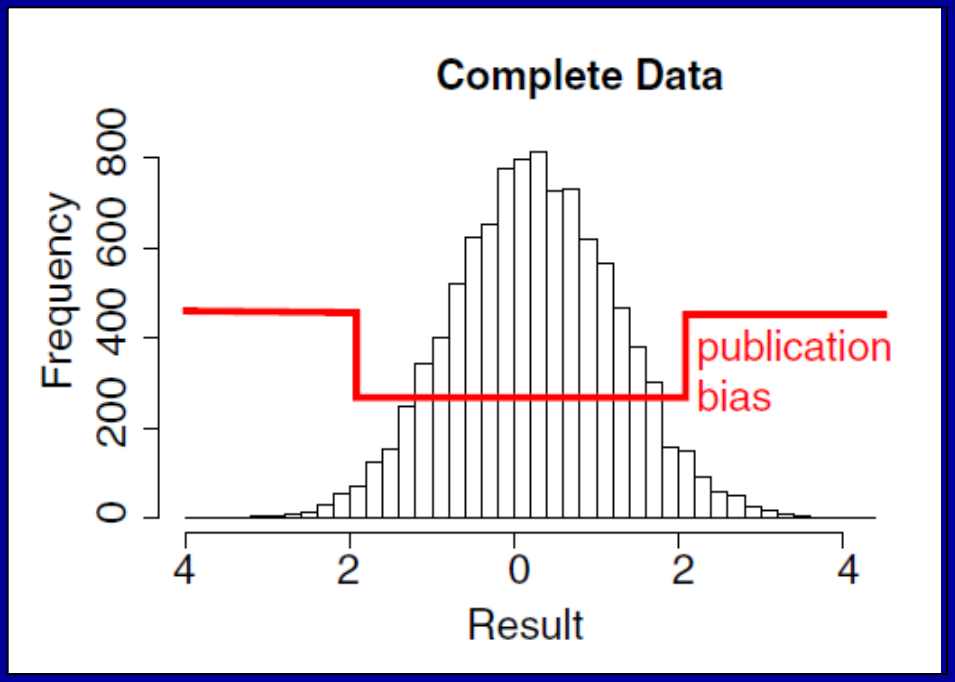
Figure 1. *P*-curves for different true effect sizes in the presence and absence of *p*-hacking. Graphs depict expected *p*-curves for difference-of-means *t* tests for samples from populations with means differing by *d* standard deviations. A–D: These graphs are products of the central and noncentral *t* distribution (see Supplemental Material 1). E–H: These graphs are products of 400,000 simulations of two samples with 20 normally distributed observations. For 1E–1H, if the difference was not significant, five additional, independent observations were added to each sample, up to a maximum of 40 observations. Share of $p < .05$ indicates the share of all studies producing a statistically significant effect using a two-tailed test for a directional prediction (hence 2.5% under the null).

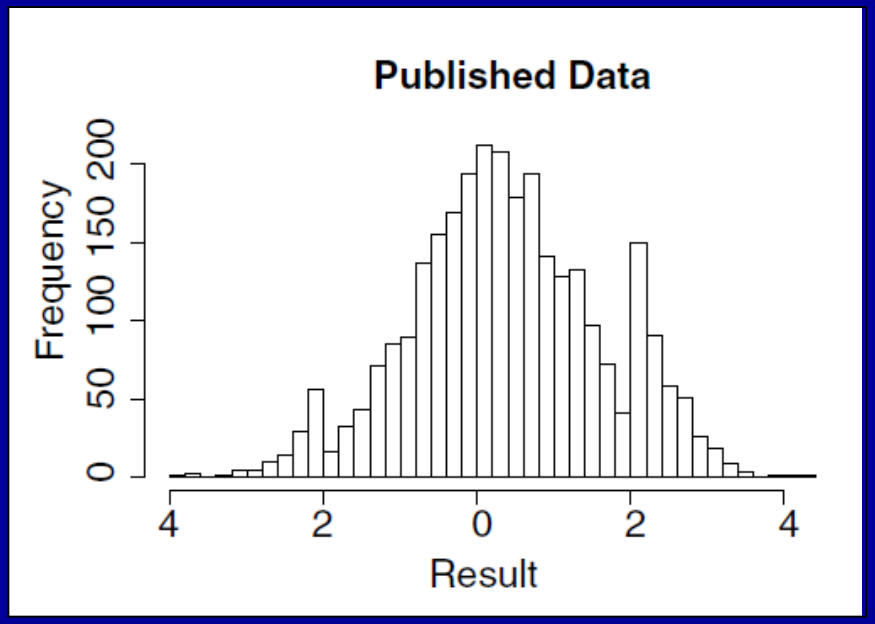
Extremely tiny bias can cause p-curves that falsely resemble genuine effects



Bruns and Ioannidis, PLoS ONE 2016

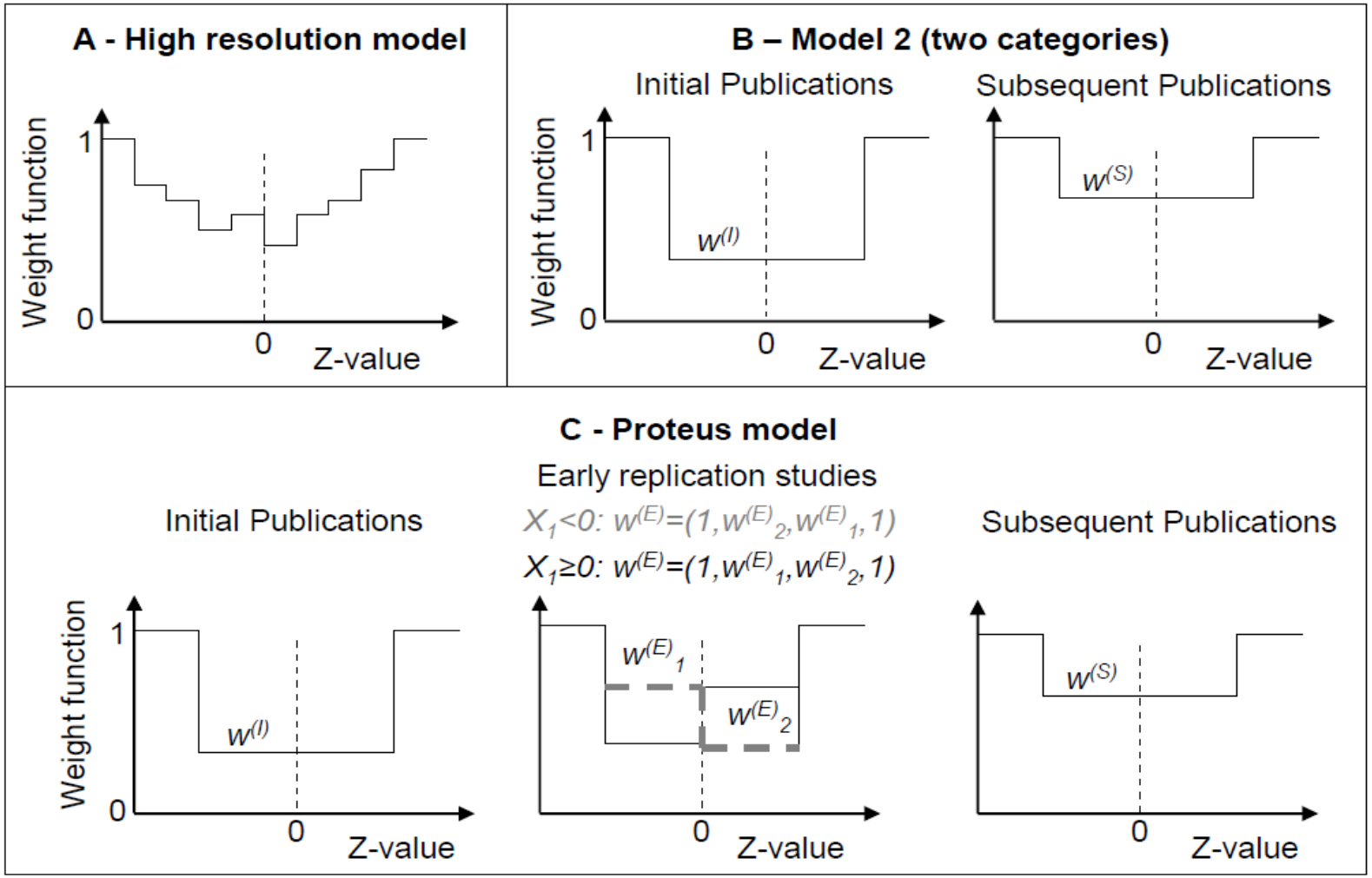
Modeling the publication selection process





Pfeiffer, Bertram, Ioannidis, PLoS ONE 2011

Adding different selection processes for initial studies, early replications, late replications



Selecting the selection model for p-values: publication bias, early effects, and Proteus phenomenon

	Random-effects model				
	Unbiased	**Model 1**	**Model 2**	**Model 3**	**Proteus**
$\log w^{(I)}$	-	-	-0.81 (0.17)	-0.82 (0.17)	-0.81 (0.17)
$\log w^{(E)}{}_1$	-	-	-	-0.33 (0.17)	-0.11 (0.17)
$\log w^{(E)}{}_2$	-	-	-	-0.24 (0.17)	-0.43 (0.17)
$\log w^{(S)}$	-	-0.33 (0.11)	-0.17 (0.12)	-0.08 (0.14)	-0.08 (0.14)
Δ_L	0	4.4	10.6	11.4	13.9
Parameters	0	1	2	4	4
Δ_{AIC}	0	-6.8	-17.2	-14.8	-19.8

Pfeiffer, Bertram, Ioannidis, PLoS ONE 2011

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Evaluation of Excess Significance Bias in Animal Studies of Neurological Diseases

Konstantinos K. Tsilidis[1☯], Orestis A. Panagiotou[1☯], Emily S. Sena[2,3], Eleni Aretouli[4,5], Evangelos Evangelou[1], David W. Howells[3], Rustam Al-Shahi Salman[2], Malcolm R. Macleod[2], John P. A. Ioannidis[6*]

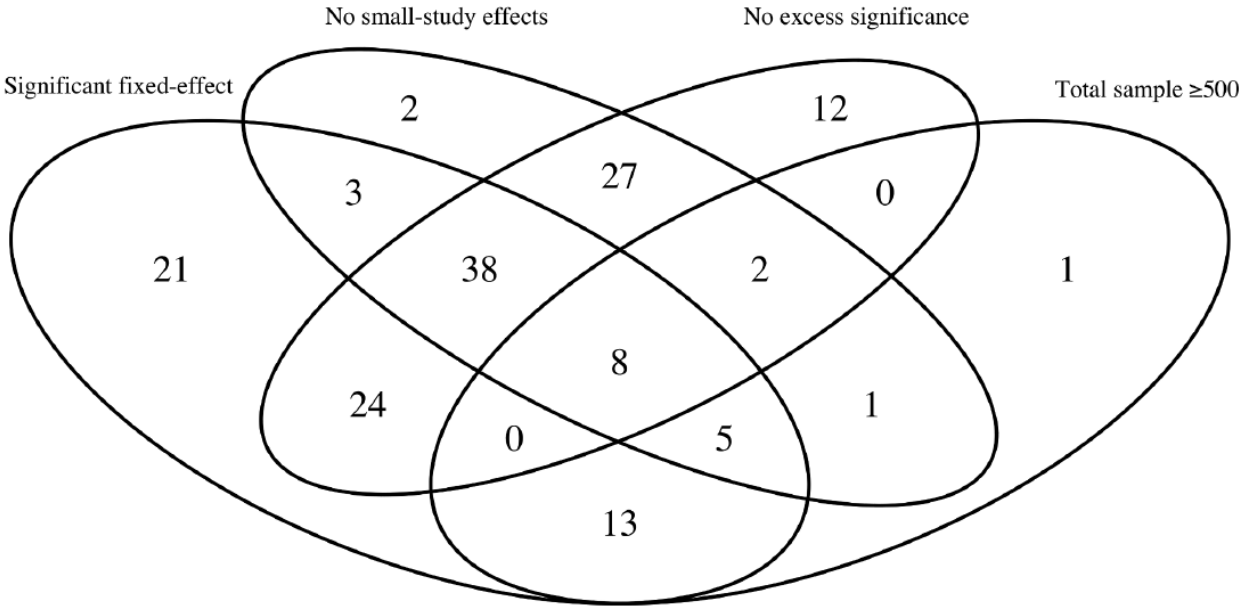
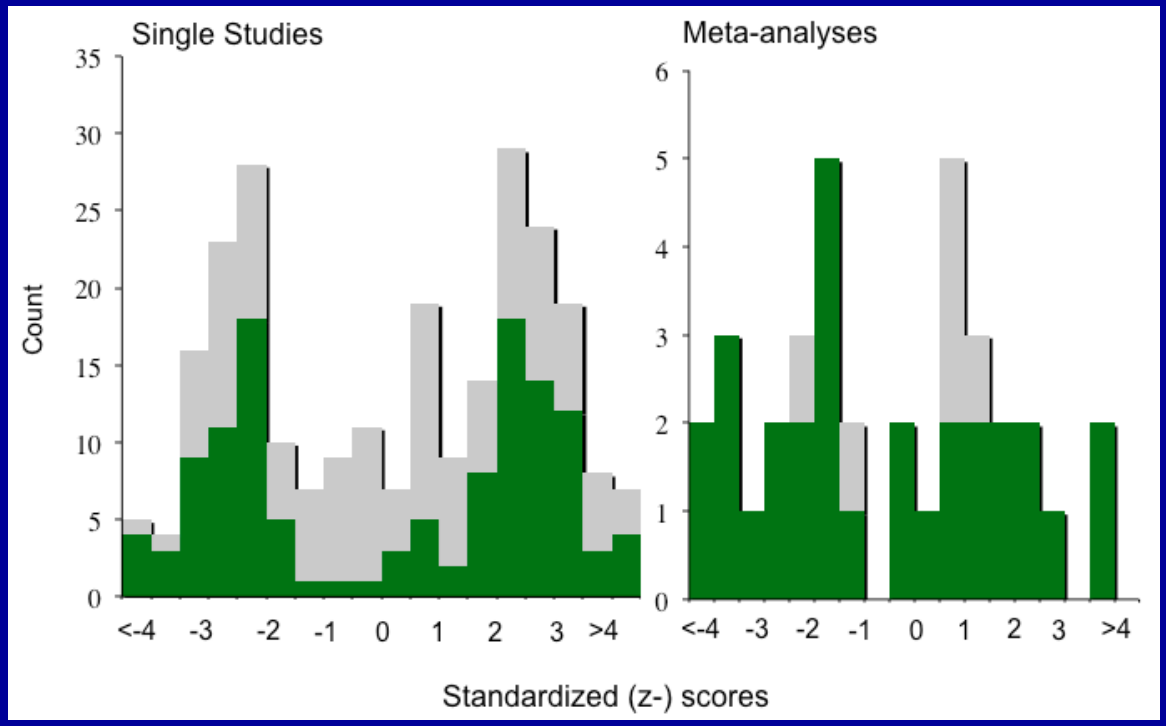


Figure 1. Venn diagrams of the meta-analyses of animal studies of neurological disorders. We plotted the number of studies with a total sample size of at least 500 animals; those which showed a nominally ($p \leq 0.05$) statistically significant effect per fixed-effects synthesis; those that had no evidence of small-study effects; and those that had no evidence of excess significance. The numbers represent the studies that have two or more of the above characteristics according to the respective overlapping areas.
doi:10.1371/journal.pbio.1001609.g001

Meta-analyses can fix only a small part of the problem



Schoenfeld and Ioannidis, AJCN 2013

Problems with methods that have inappropriate familywise error rates

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; [c]Center for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and [e]WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

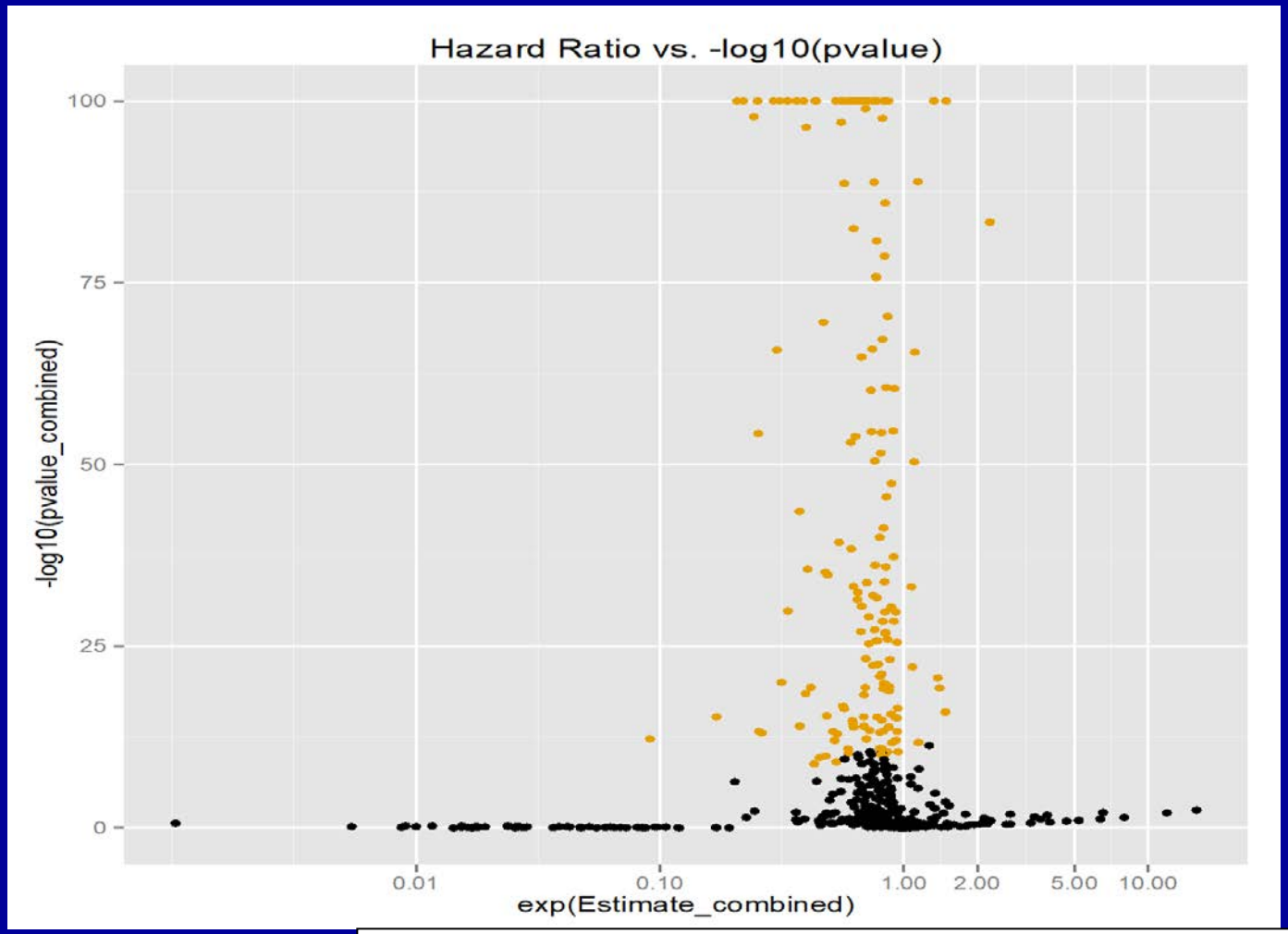
Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data and a total of 3 million random task group analyses to compute empirical familywise error rates for the fMRI software packages SPM, FSL, and AFNI, as well as a nonparametric permutation method. For a nominal familywise error rate of 5%, the parametric statistical methods are shown to be conservative for voxelwise inference and invalid for clusterwise inference. Our results suggest that the principal cause of the invalid cluster inferences is spatial autocorrelation functions that do not follow the assumed Gaussian shape. By comparison, the nonparametric permutation test is found to produce nominal results for voxelwise as well as clusterwise inference. These findings speak to the need of validating the statistical methods being used in the field of neuroimaging.

(FWE), the chance of one or more false positives, and empirically measure the FWE as the proportion of analyses that give rise to any significant results. Here, we consider both two-sample and one-sample designs. Because two groups of subjects are randomly drawn from a large group of healthy controls, the null hypothesis of no group difference in brain activation should be true. Moreover, because the resting-state fMRI data should contain no consistent shifts in blood oxygen level-dependent (BOLD) activity, for a single group of subjects the null hypothesis of mean zero activation should also be true. We evaluate FWE control for both voxelwise inference, where significance is individually assessed at each voxel, and clusterwise inference (19–21), where significance is assessed on clusters formed with an arbitrary threshold.

In brief, we find that all three packages have conservative voxelwise inference and invalid clusterwise inference, for both one- and two-sample t tests. Alarmingly, the parametric methods

Problems with big data: One third of known medications may affect cancer risk (!?)



Patel et al, Sci Rep 2016

¾ of medication classes may affect cancer risk (!?)

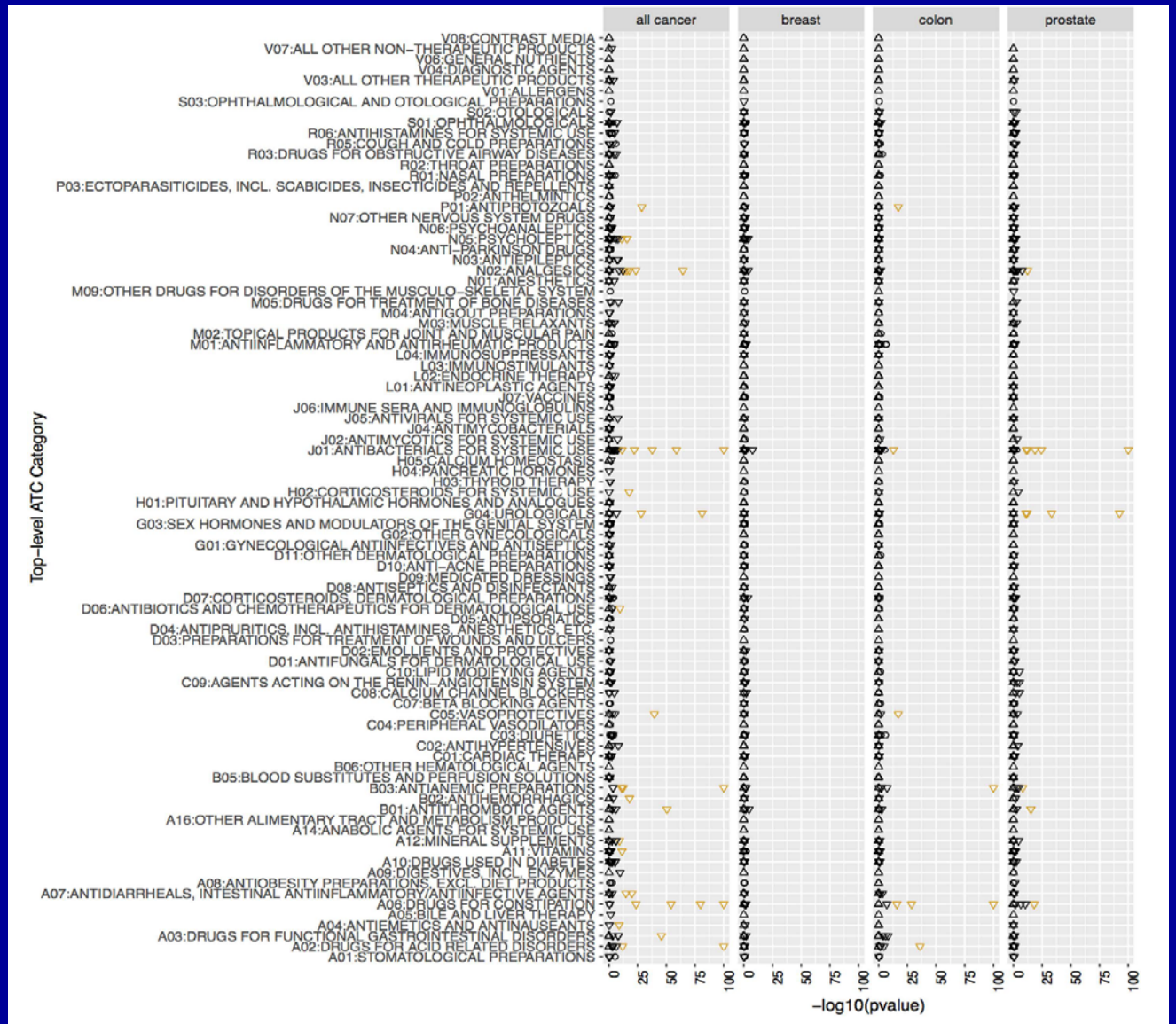
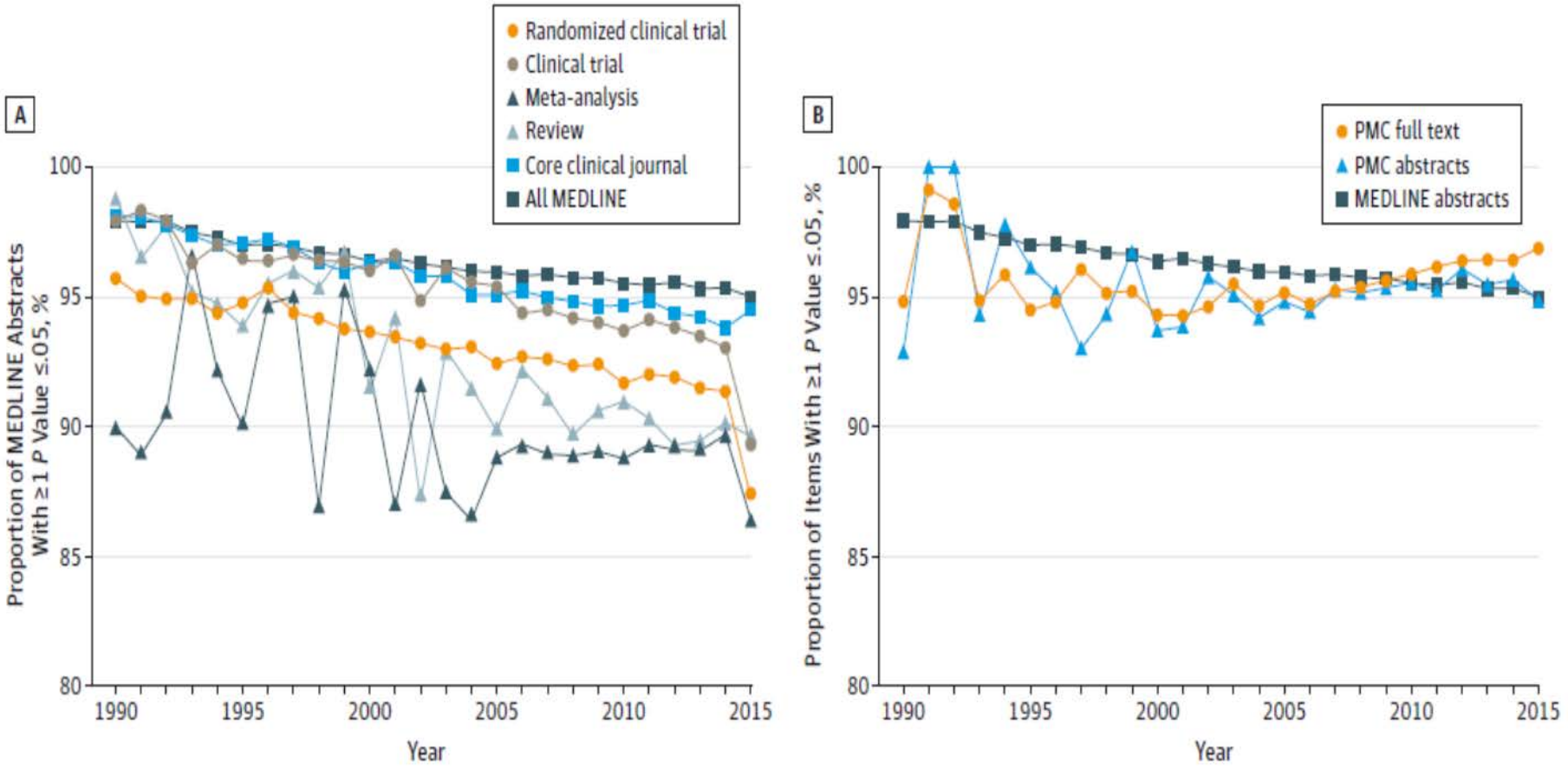


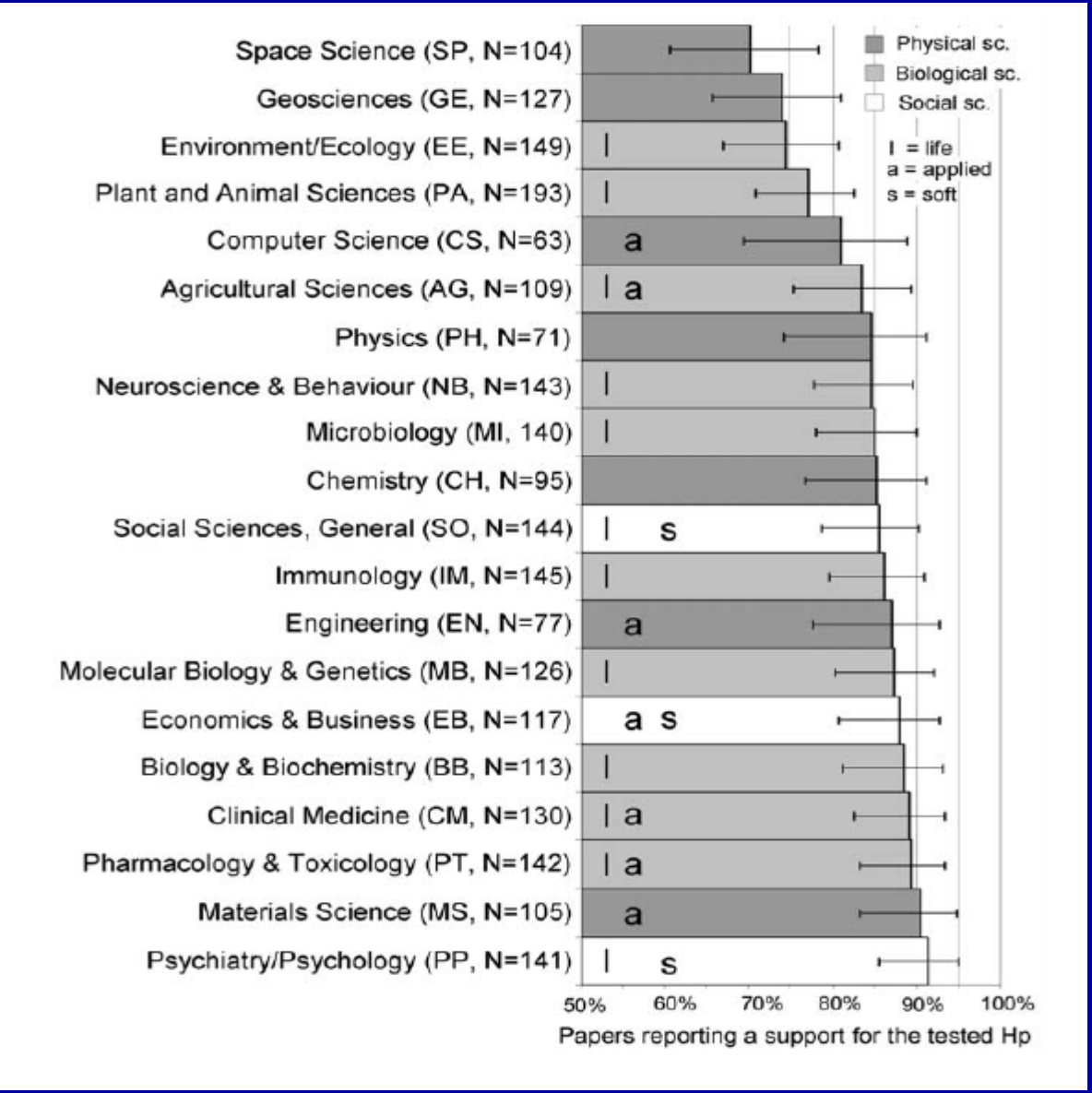
Figure 3. Manhattan plot (−log10(p-value)) for each drug categorized by Anatomical Therapeutic Chemical anatomical main group) in case-crossover analyses. Orange color denotes tentative signals. P-values lower than 1×10^{-100} set to 1×10^{-100} for clarity. Upward triangles indicate OR >1 and downward triangles OR <1.

Statistical significance has become a boring nuisance: 96% of the biomedical literature claims significant results (and the vast majority of them say they are novel)



Chavalarias, Wallach, Li, Ioannidis, JAMA 2016

Fields with the highest proportion of statistically significant claims may be the least reliable



Fanelli, PLoS ONE 2010

In-depth assessment of 1000 abstracts

	%
P – values	
Abstracts with at least one P-value	12.5
Abstracts with at least one statistically significant P-value (≤ 0.05)	11.8
Abstracts with at least one statistically non significant P-value (> 0.05)	2.7
Confidence Intervals (CI)	
Abstracts with at least one CI	1.8
Abstracts with at least one 95% CI	1.7
Effect sizes	
Abstract with at least one effect size	11.0
Abstracts with at least one effect size and one P-value for at least one of the effect size	3.7
Abstracts with at least one effect size and one 95% CI for at least one effect size	1.4
Effect sizes that can be calculated	
Abstracts where at least one effect size can be calculated	9.9
Abstracts where at least one effect size can be calculated and one P-value is reported for at least one effect size that can be calculated	4.2
Abstracts with at least one effect size that can be calculated and one CI is reported for at least one effect size that can be calculated	0.0
Qualitative statements about significance	
Abstracts with at least one statement about significance	18.2
Abstracts with at least one statement about significance, with at least one effect size or where one effect size can be calculated	2.6
Abstracts with at least one statement about significance, with at least one effect size or where one effect size can be calculated and at least one confidence interval for at least one effect size	0.3

Effect sizes and effect sizes that can be calculated

Effect size	Total n = 264	With P –value	With CI	With P-value or CI
Relative risk / Risk ratio	8	5 (62.5)	3 (37.5)	8 (100.0)
Percent difference/change	78	10 (12.8)	6 (7.7)	16 (100.0)
Mean difference/change	4	0 (0.0)	2 (50.0)	2 (50.0)
Correlation coefficient	58	18 (31.0)	0 (0.0)	18 (31.0)
Absolute difference/change	29	10 (34.5)	4 (13.8)	14 (48.3)
Beta coefficient	8	3 (3.8)	0 (0.0)	3 (3.8)
Odds ratio	15	8 (53.3)	12 (80.0)	14 (93.3))
Hazard ratio	12	5 (41.7)	7 (58.3)	12 (100.0)
Fold difference/change	34	1 (2.9)	0 (0.0)	1 (2.9)
Interclass correlation coefficient	4	0 (0.0)	0 (0.0)	0 (0.0)
Relative risk reduction	6	0 (0.0)	6 (100.0)	6 (100.0)
Assorted ratio	8	5 (62.5)	2 (25.0)	5 (62.5)
Where effect sizes could be calculated	Total n = 221	With P –value	With CI	With P-value or CI
Comparison of means	20	5 (25.0)	0 (0.0)	5 (25.0)
Absolute comparisons	60	27 (45.0)	0 (0.0)	27 (45.0)
Comparison of medians	7	6 (85.7)	0 (0.0)	6 (85.7)
Comparison of proportions	134	41 (30.6)	0 (0.0)	41 (30.6)

Among 100 full-text articles from PubMed

- 55 report P-values
- 4 present CIs for all the reported effect sizes
- none use Bayesian methods
- none use false-discovery rate methods

Two trials with p<0.05 (FDA rule) – what does it mean in BF terms?

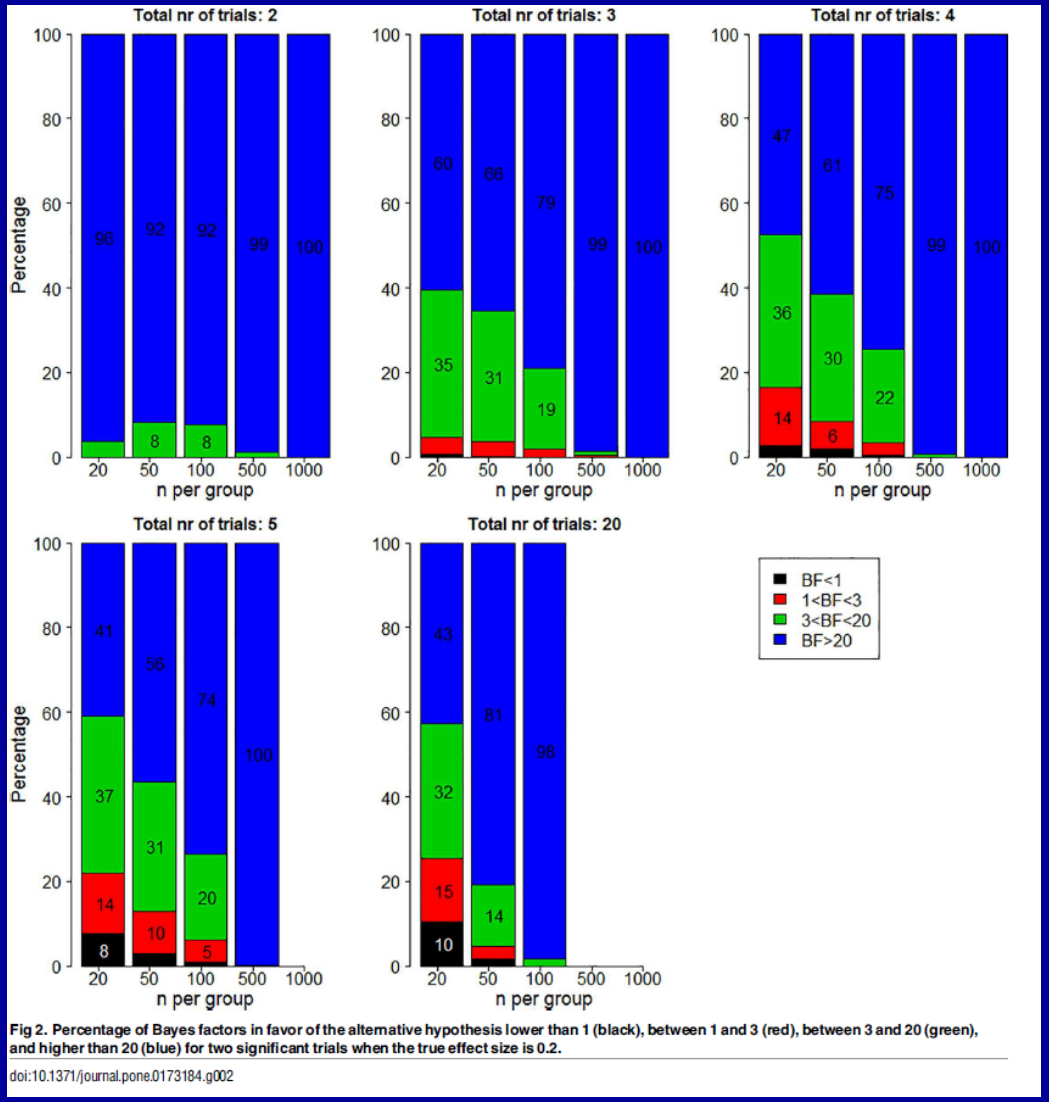


Fig 2. Percentage of Bayes factors in favor of the alternative hypothesis lower than 1 (black), between 1 and 3 (red), between 3 and 20 (green), and higher than 20 (blue) for two significant trials when the true effect size is 0.2.

doi:10.1371/journal.pone.0173184.g002

Van Ravenzwaaij and Ioannidis
PLoS ONE 2017

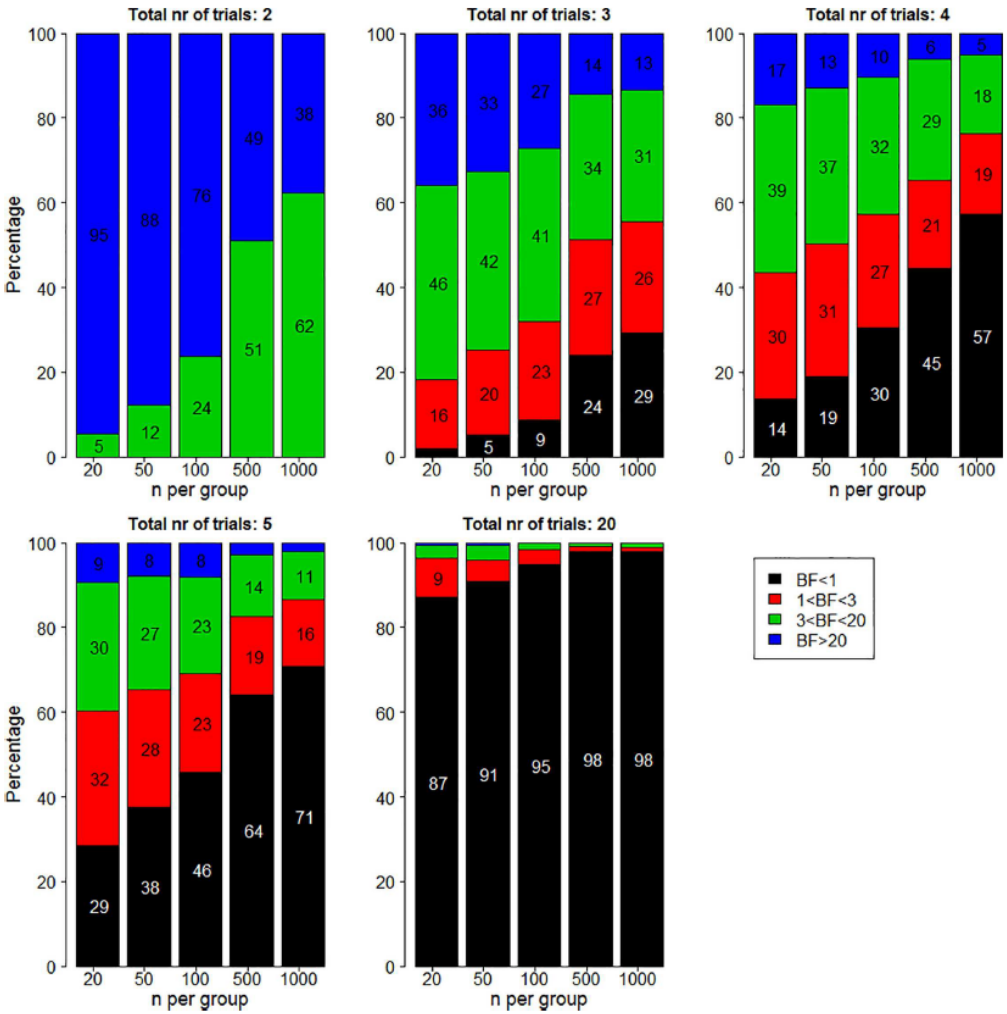


Fig 6. Percentage of Bayes factors in favor of the alternative hypothesis lower than 1 (left panel), lower than 3 (middle panel), and lower than 20 (right panel) for two significant trials when the true effect size is 0.

doi:10.1371/journal.pone.0173184.g006

Redefine statistical significance

Nature Human Behavior
September 1, 2017

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive. While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new[1,2], a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (for example, genomics and high-energy physics research; see the 'Potential objections' section below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P* values. However, changing the *P* value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P* values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P* value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the *P* value falls below the (current) type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{obs})}{\Pr(H_0 \mid x_{obs})} = \frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \equiv \mathrm{BF} \times (\text{prior odds}) \qquad (1)$$

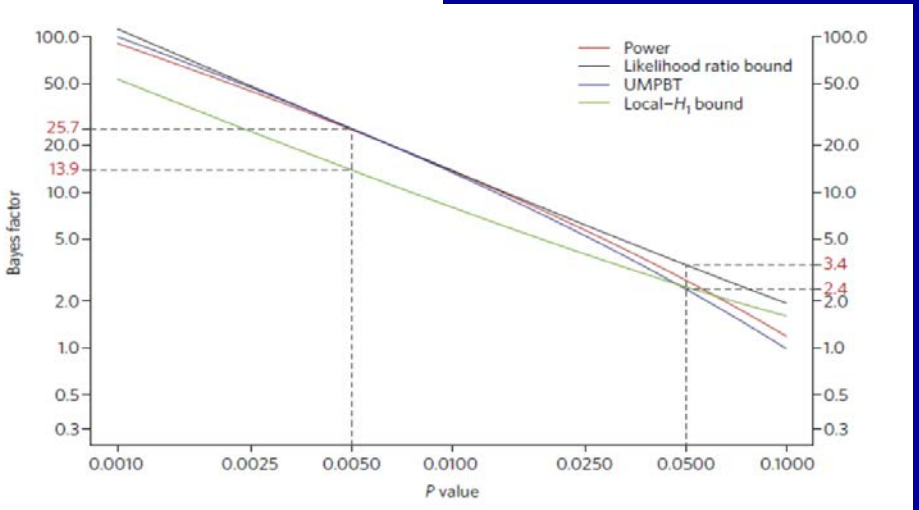


Fig. 1 | Relationship between the *P* value and the Bayes factor. The Bayes factor (BF) is defined as $\frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)}$. The figure assumes that observations are independent and identically distributed

frontiers in Human Neuroscience

REVIEW
published: 03 August 2017
doi: 10.3389/fnhum.2017.00390

When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment

Denes Szucs[1]* *and John P. A. Ioannidis*[2]

[1] *Department of Psychology, University of Cambridge, Cambridge, United Kingdom,* [2] *Meta-Research Innovation Center at Stanford and Department of Medicine, Department of Health Research and Policy, and Department of Statistics, Stanford University, Stanford, CA, United States*

Null hypothesis significance testing (NHST) has several shortcomings that are likely contributing factors behind the widely debated replication crisis of (cognitive) neuroscience, psychology, and biomedical science in general. We review these shortcomings and suggest that, after sustained negative experience, NHST should no longer be the default, dominant statistical practice of all biomedical and psychological research. If theoretical predictions are weak we should not rely on all or nothing hypothesis tests. Different inferential methods may be most suitable for different types of research questions. Whenever researchers use NHST they should justify its use, and publish pre-study power calculations and effect sizes, including negative findings. Hypothesis-testing studies should be pre-registered and optimally raw data published. The current statistics lite educational approach for students that has sustained the widespread, spurious use of NHST should be phased out.

OPEN ACCESS

Edited by:

Is NHST a good choice for:

- Developing a prognostic score for cardiovascular disease?
- Assessing a diagnostic test for depression?
- Evaluating a medical therapy in a randomized trial?
- Mining electronic health records?
- Mining big data from metabolomics?
- Assessing if women athletes with high natural testosterone should be excluded from the Olympics?

Concluding comments

- The use of P-values has become an epidemic affecting the majority of scientific disciplines
- Strong selection biases make almost everything (seem) statistically significant
- NHST and P-values are inherently most suitable/optimal for only a minority of current research
- Using a more stringent threshold is a temporizing measure to avoid death-by-significance
- NHST and P-values may be replaced in many fields by other inferential methods
- Selection biases will need more drastic measures to be curtailed rather than just a change in inferential method

Special thanks

- David Chalavarias, ICS, Paris
- Steve Goodman, Stanford
- Dan Fanelli, Stanford and LSE
- Josh Wallach, Stanford and Yale
- Denes Szucs, Cambridge
- Chirag Patel, Harvard and Stanford
- Mark Cullen, Stanford
- David Rehkopf, Stanford
- Ioanna Tzoulaki, Imperial College
- Athina Tatsioni, U Ioannina
- Nikos Patsopoulos, Harvard
- Kevin Boyack, SciTech
- Stephan Bruns, U Kassel
- Belinda Burford, U Melbourne
- Jonathan Schoenfeld, Harvard
- Shanil Ebrahim, McMaster U
- Muin Khoury, CDC and NCI/NIH
- Stelios Serghiou, Stanford
- Muin Khoury, NIH and CDC
- Sheri Schully, NIH, NCI
- Kostas Tsilidis, U Ioannina
- Alvin Li, U Western Ontario
- Dan van Ravenzwaaij, U Utrecht
- Ioana Cristea, Stanford and U Cluj