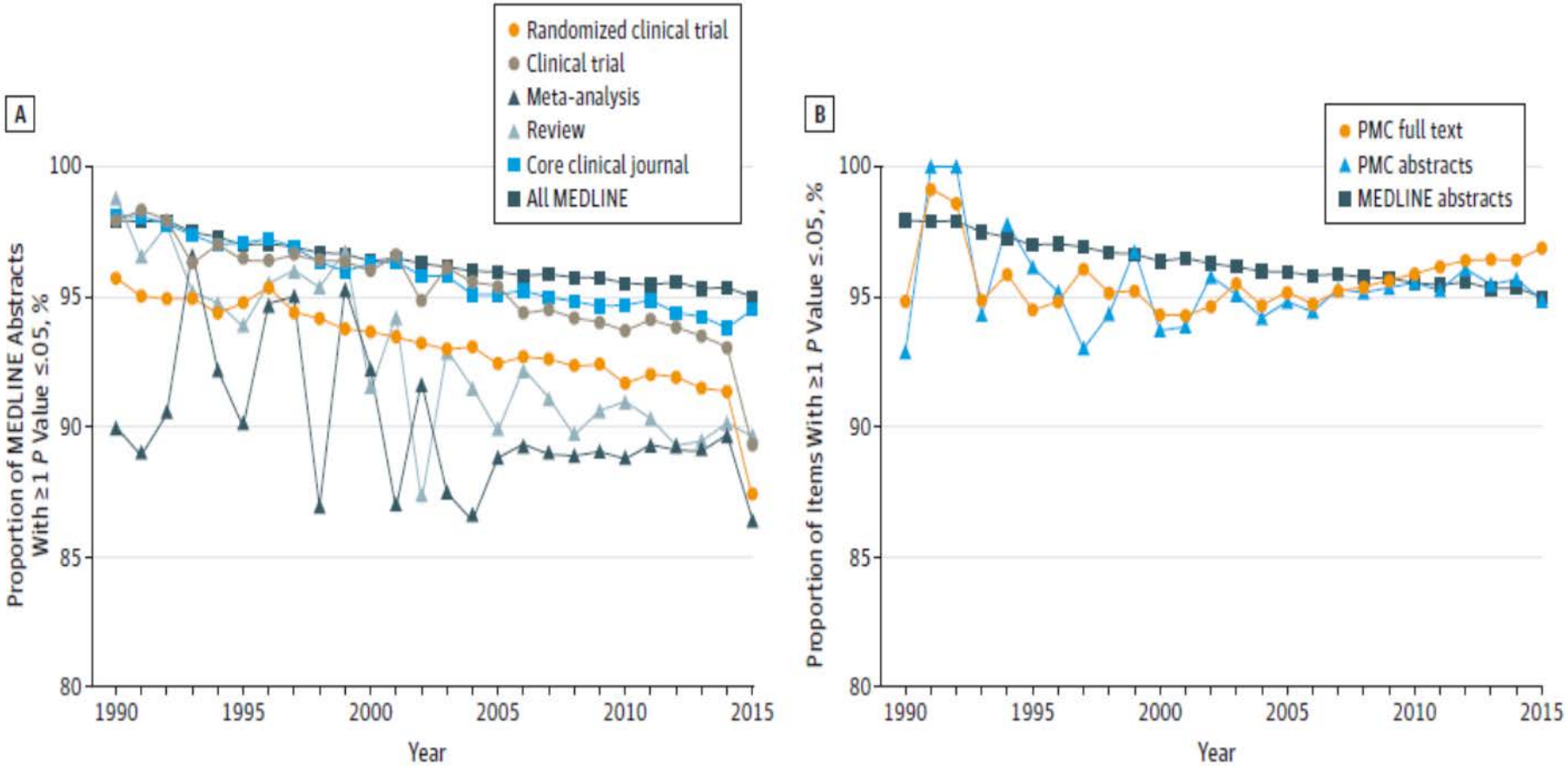
The reproducibility crisis in academic research

Karolinska, 10/2019

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
Co-Director, Meta-Research Innovation Center at Stanford (METRICS)
Director, PhD program in epidemiology and clinical research
Stanford University
Visiting Einstein fellow, Berlin Institute of Health

Scientific discovery has become a boring nuisance: 96% of the scientific literature claims significant results



Chavalarias, Wallach, Li, Ioannidis, JAMA 2016

Discovery and/or replication: what value?

- Let *R*=pre-study odds for a research finding, BF=Bayes factor conferred by the discovery data, *h*=ratio of the weight of negative consequences from FP discovery claim versus the positive consequences from TP discovery.
- Value of the discovery is proportional to TP – (*h* * FP) or (TP/FP) – *h*=(*R* * BF) – *h*.
- *R* and *h* are rather field specific and cannot be modified (unless you change field).
- Focus must be on increasing BF

Ioannidis. Behavioral and Brain Sciences 2018, 41:e137

Many/most “discoveries” may have negative scientific value unless replicated

- Options for increasing BF: running larger studies and ensuring greater protection from biases. And, of course, replication.
- To avoid negative values for the value of discovery, one needs BF>h/R. Often this is difficult in the absence of replication.
- Most original discoveries come from small studies, where biases are common, BF is <5 and R is very low.

Ioannidis. Behavioral and Brain Sciences 2018, 41:e137

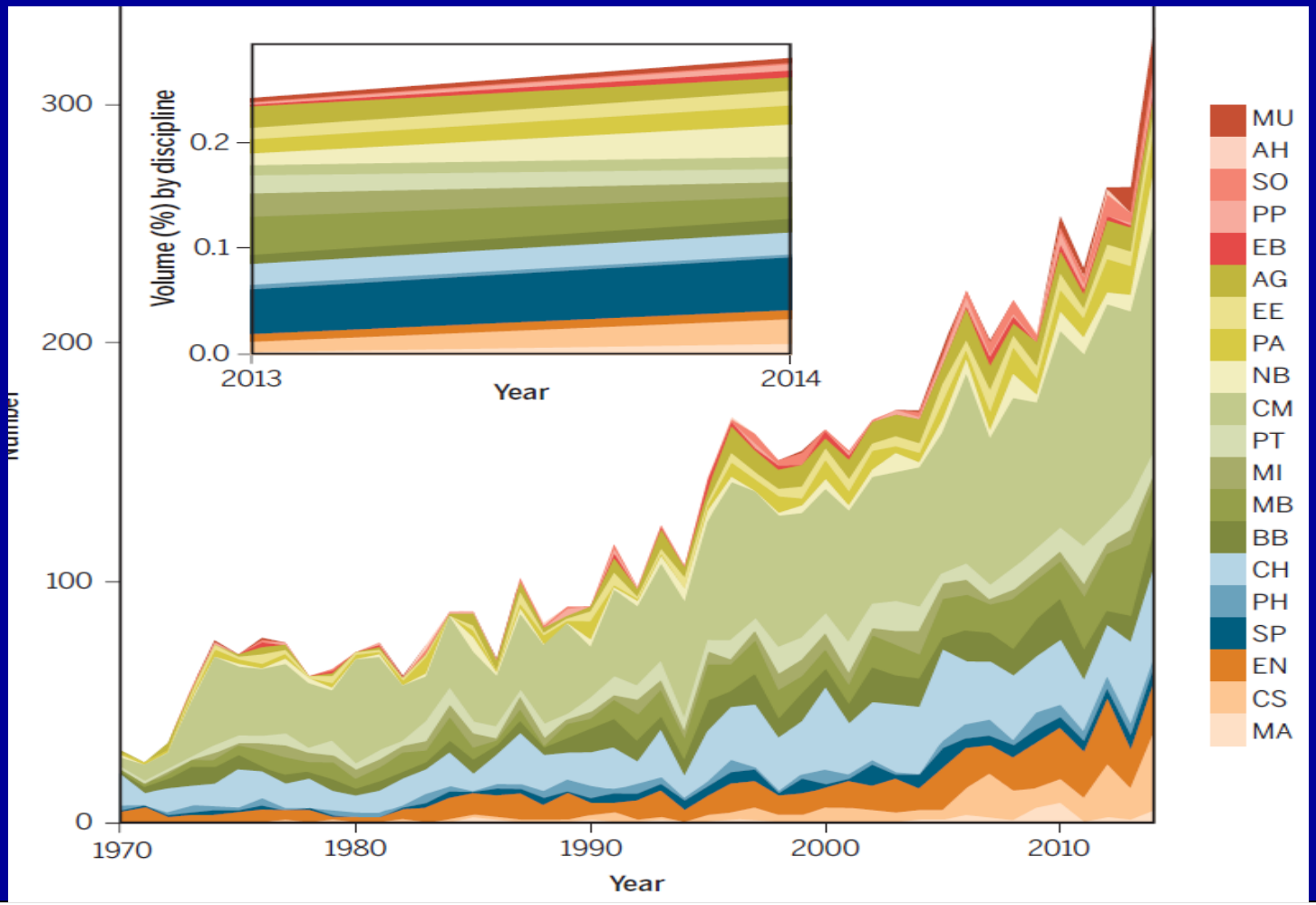
Reproducibility
Confusing terminology

- Reproducibility
- Replication
  - Exact replication
  - Conceptual replication
- Re-analysis
- Repeatability
- Corroboration
- Triangulation

What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of "research reproducibility" are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for "truth."



Goodman, Fanelli, Ioannidis. Science Translational Medicine 2016

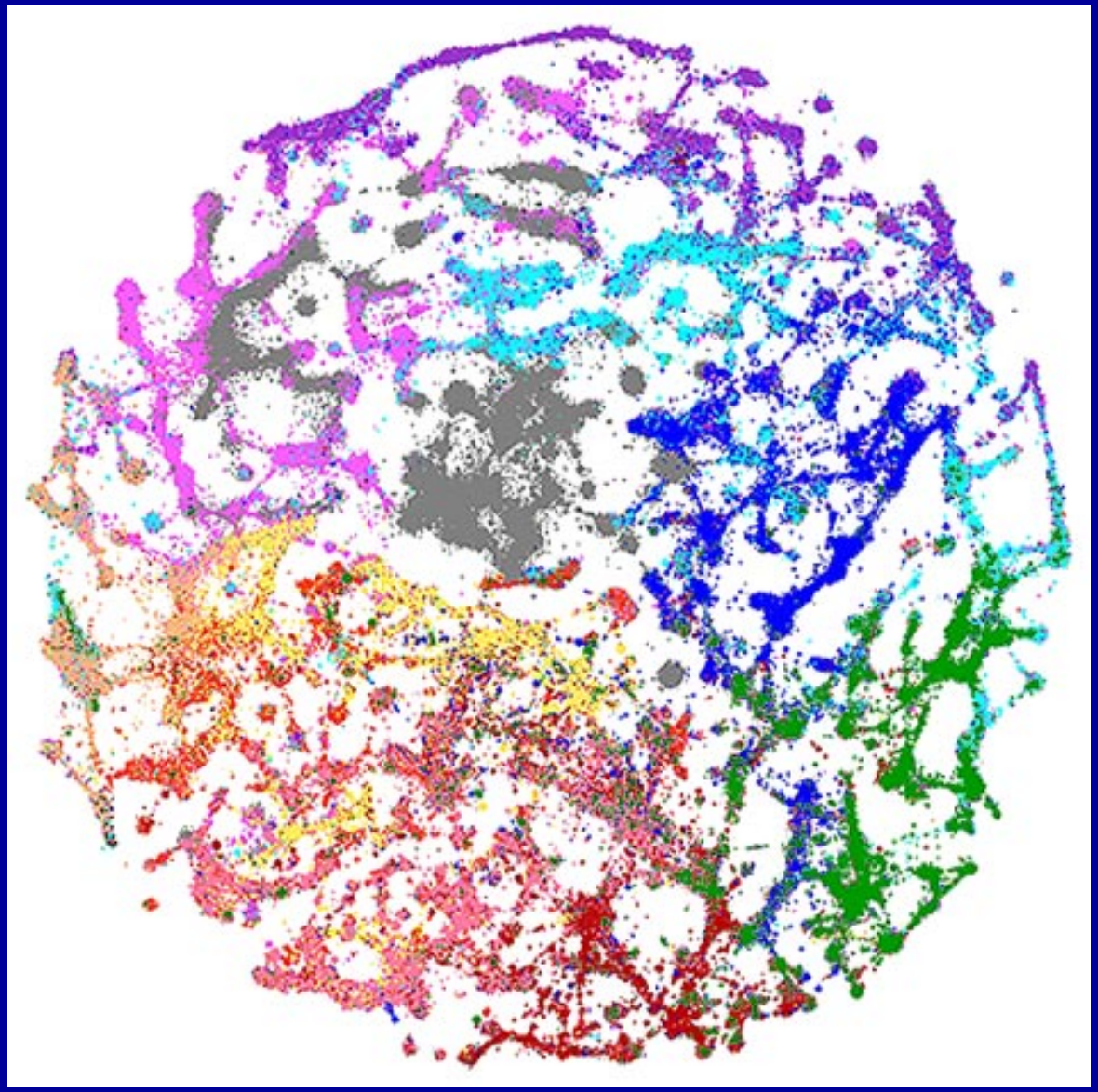
Different types of reproducibility

- **Reproducibility of methods:** the ability to understand or repeat as exactly as possible the experimental and computational procedures.
- **Reproducibility of results:** the ability to produce corroborating results in a new study, having followed the same experimental methods.
- **Reproducibility of inferences:** the making of knowledge claims of similar strength from some study results.

Differences across fields that affect what "reproducible research" means

- Degree of determinism
- Signal to measurement-error ratio
- Complexity of designs/measurement tools
- Closeness of fit between hypothesis and experimental design/data.
- Statistical/analytic methods to test hypotheses
- Typical heterogeneity of experimental results
- Culture of replication, transparency and cumulating knowledge
- Statistical criteria for truth claims
- Purposes to which findings will be put and consequences of false conclusions.

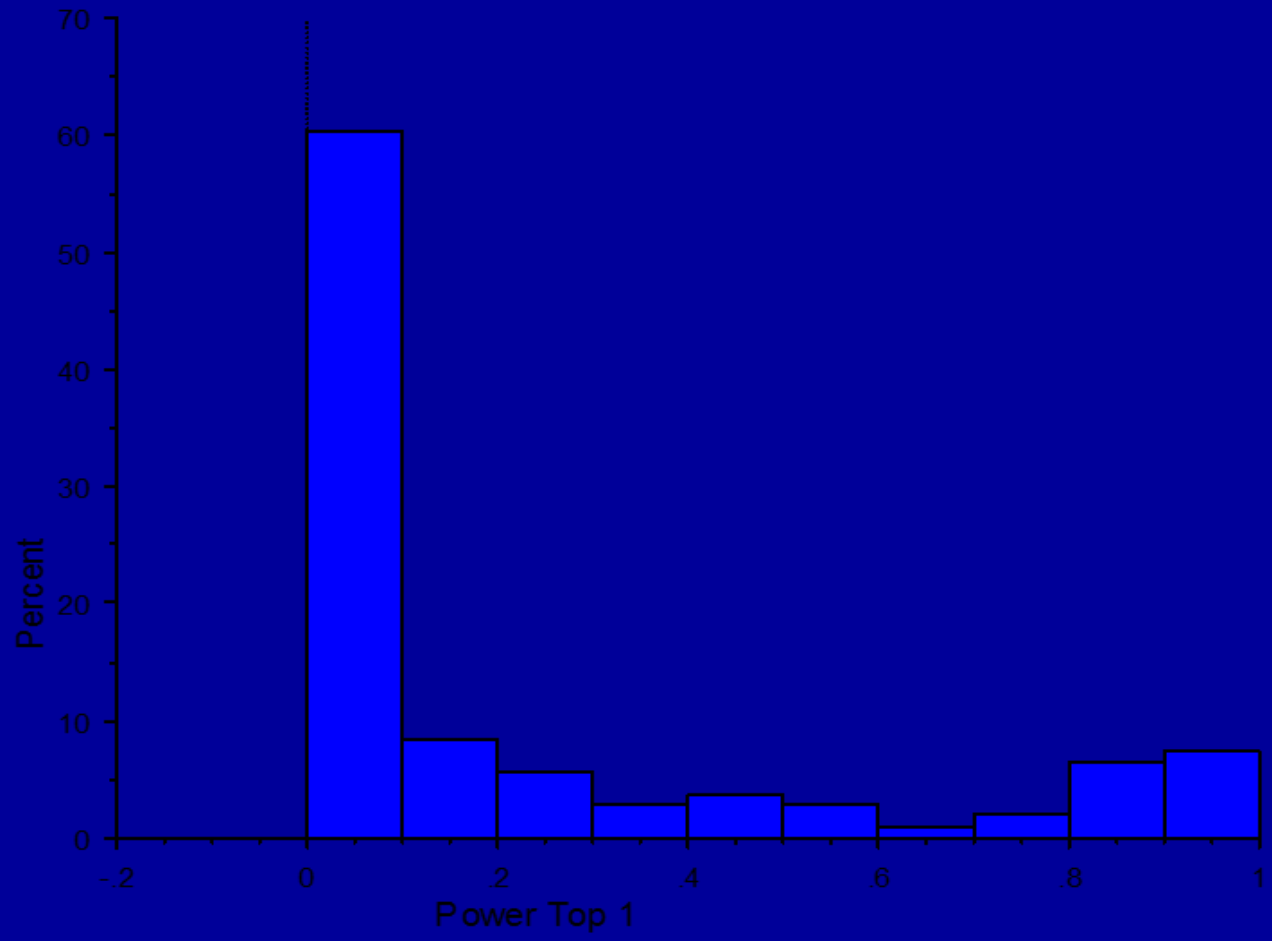
A map of scientific research



Typical recipe of research practices: small data

- Small sample size studies
- Solo, siloed investigator, small team
- Cherry-picking of one/best hypothesis
- Post-hoc
- $P<0.05$ is enough
- No registration
- No data sharing
- No replication

Power in 130 economics topics (>10,000 studies with >70,000 effect estimates)



Ioannidis, Stanly, Doucouliagos, The Economic Journal 2017

Typical recipe of research practices: big data

- Extremely large sample size (overpowered) studies
- Cherry-picking of one/best hypothesis
- Post-hoc
- Idiosyncratic statistical inference tools without consensus
- No registration
- Data sharing without understanding what is shared

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

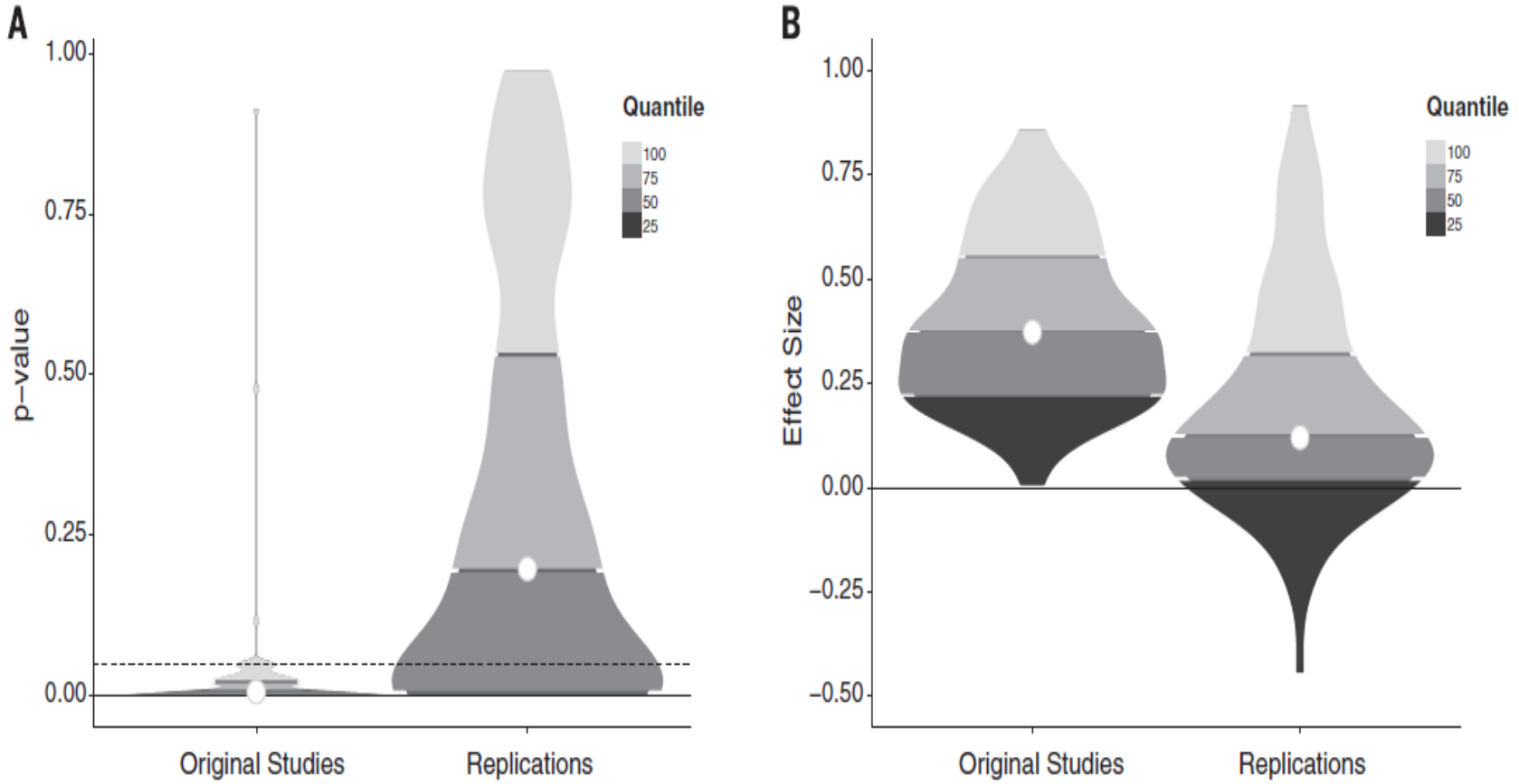
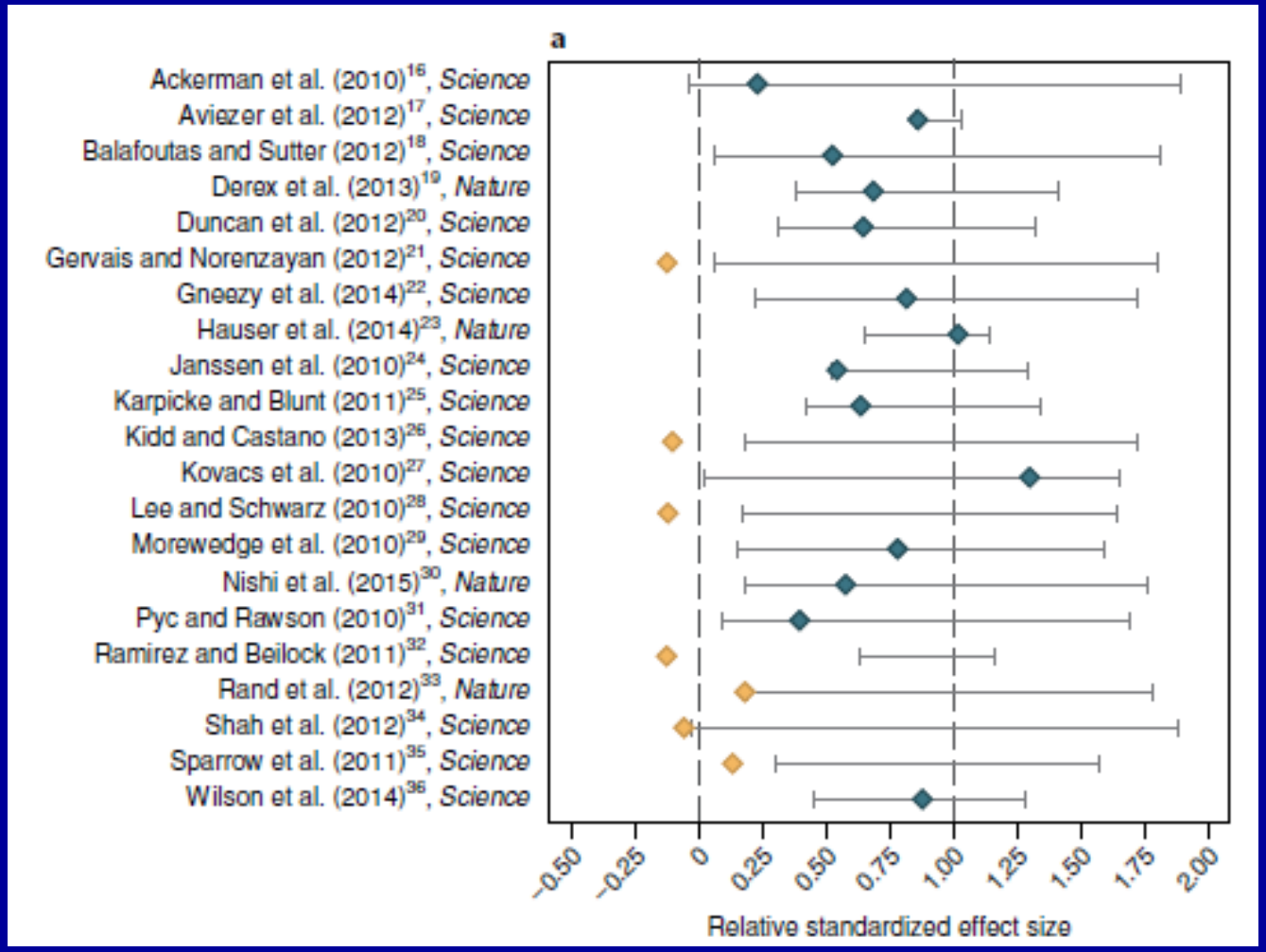


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

What if I only read Nature and Science?



Camerer et al. Nature HB 2018

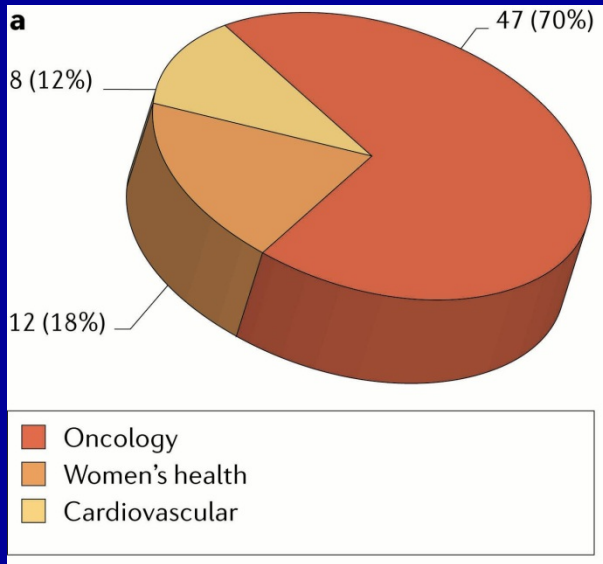
Candidate genes replicated through GWAS: replication rate = 1.2%

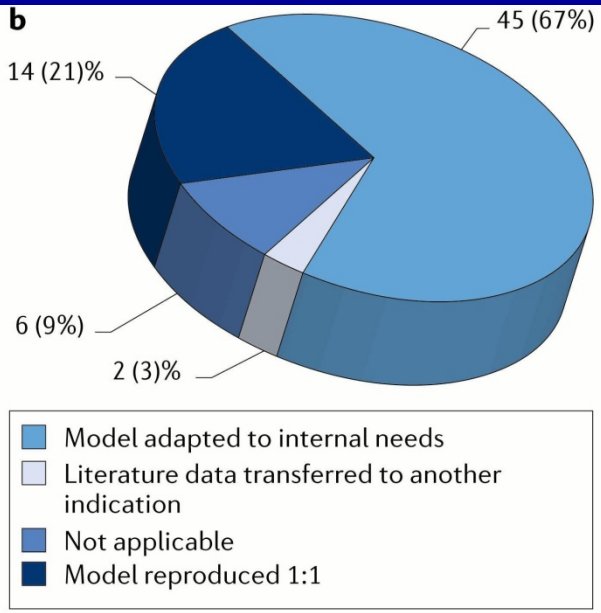
Table. Large-scale efforts to massively replicate reported candidate gene associations

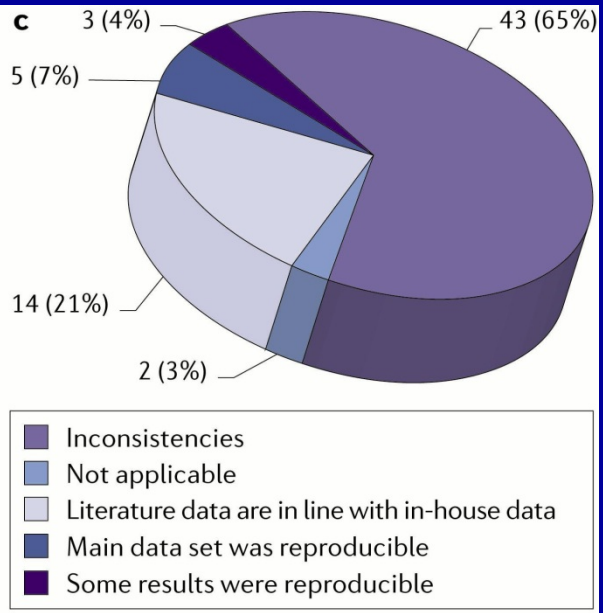
First author	Disease/phenotype	Gene loci tested	Sample size (design)	Replicated gene loci
Bosker (16)	Major depressive disorder	57	3540 (Case-control)	1
Caporaso (17)	Smoking (7 phenotypes)	359	4611 (Cohort)	1
Morgan (18)	Acute coronary syndrome	70	1461 (Case-control)	0
Richards (19)	Osteoporosis (2 phenotypes)	150	19,195 (Cohort)	9
Samani (20)	Coronary artery disease	55	4864; 2519 (Case-control)	1
Scuteri (21)	Obesity (3 phenotypes)	74	6148 (Cohort)	0
Soeber (22)	Blood pressure	149	1644; 8023 (Cohort)	0
Wu (23)	Childhood asthma	237	1476 (Triads)	1

Ioannidis, Tarone, McLaughlin, Epidemiology 2011

Failed replication in preclinical research







d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

Nature Reviews | Drug Discovery

Prinz et al., Nature Reviews Drug Discovery 2011

Replicated: only 6 of 53 landmark studies for Amgen oncology drug target projects

- “The failure to win “the war on cancer” has been blamed on many factors, … But recently a new culprit has emerged: too many basic scientific discoveries… are wrong.”

Begley et al. Nature 2012

Clinical Chemistry 63:5
000–000 (2017)

Perspectives

The Reproducibility Wars: Successful, Unsuccessful, Uninterpretable, Exact, Conceptual, Triangulated, Contested Replication

John P.A. Ioannidis[1,2,3,4*]

Meta-assessment of bias in science

Daniele Fanelli[a,1], Rodrigo Costas[b], and John P. A. Ioannidis[a,c,d,e]

Table 1. Summary of each bias pattern or risk factor for bias that was tested in our study, parameters used to test these hypotheses via meta-regression, predicted direction of the association of these parameters with effect size, and overall assessment of results obtained

Hypothesis type	Hypothesis tested	Specific factor tested	Variables measured to test the hypothesis	Predicted association with effect size	Result
Postulated bias patterns	Small-study effect		Study SE	+	S
	Gray literature bias		Gray literature (any type) vs. Journal article	–	S
	Decline effect		Year order in MA	–	P
	Early extremes		Year order in MA, regressed on absolute effect size	–	N
	Citation bias		Total citations to study	+	S
	US effect		Study from author in the US vs. Any other country	+	P
	Industry bias		Studies with authors affiliated with private industry vs. Not	+	P
Postulated risk factors for bias	Pressures to publish	Country policies	Cash incentive	+	N
			Career incentive	+	N
			Institutional incentive	+	N
		Author's productivity	(First/last) author's total publications, publications per year	+	N
		Author's impact	(First/last) total citations, average citations, average normalized citations, average journal impact, % top10 journals	+	N
	Mutual control		Team size	–	S
			Countries/author, average distance between addresses	+	S
	Individual risk factors	Career level	Years in activity (first/last) author	–	S
		Gender	(First/last) author's female name	–	N
		Research integrity	(First/last) author with ≥1 retraction	+	P

Symbols indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). Conclusions as to whether results indicate that the hypothesis was fully supported (S), partially supported (P), or not supported (N) are based on main analyses as well as secondary and robustness tests, as described in the main text.

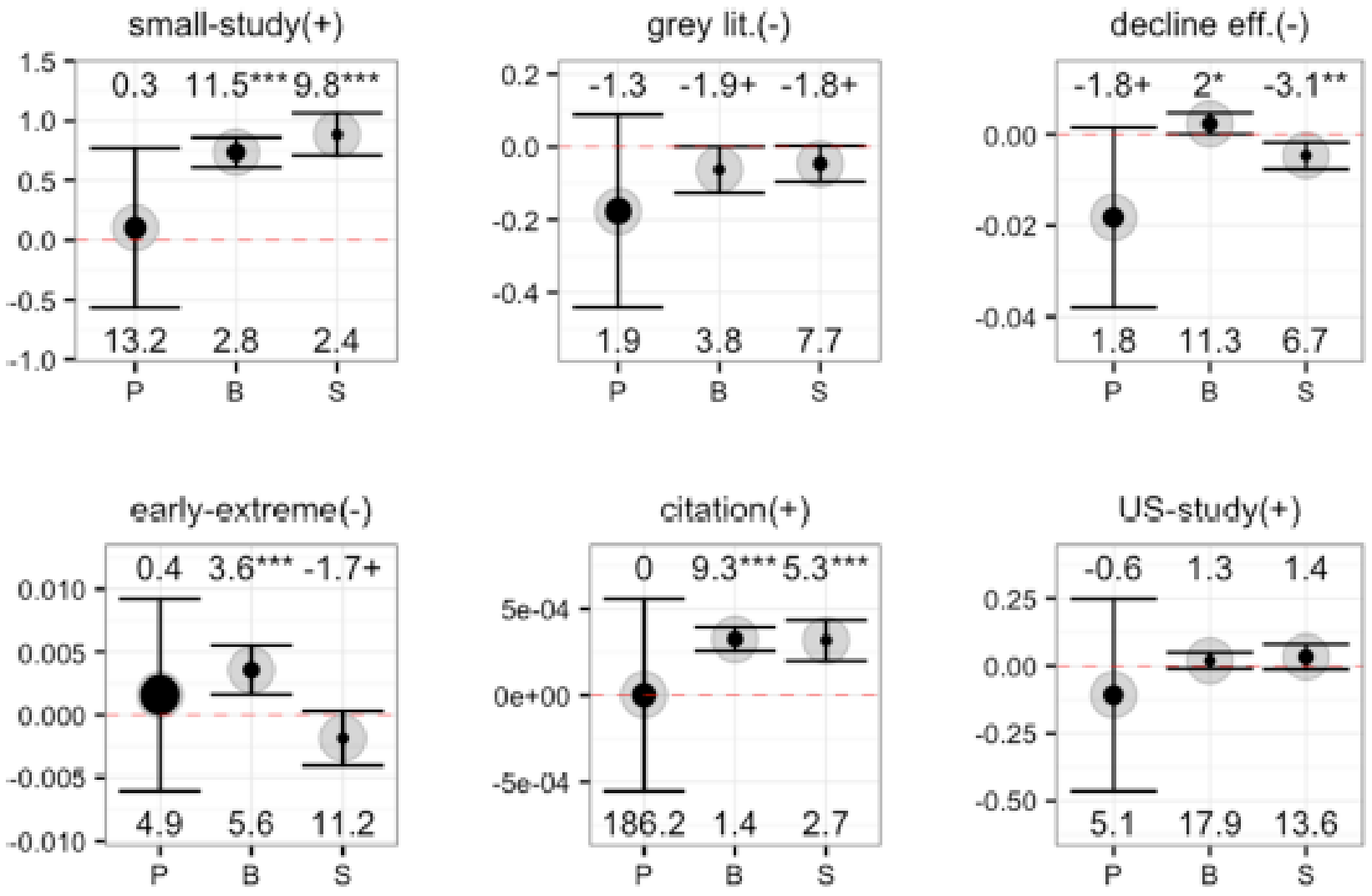


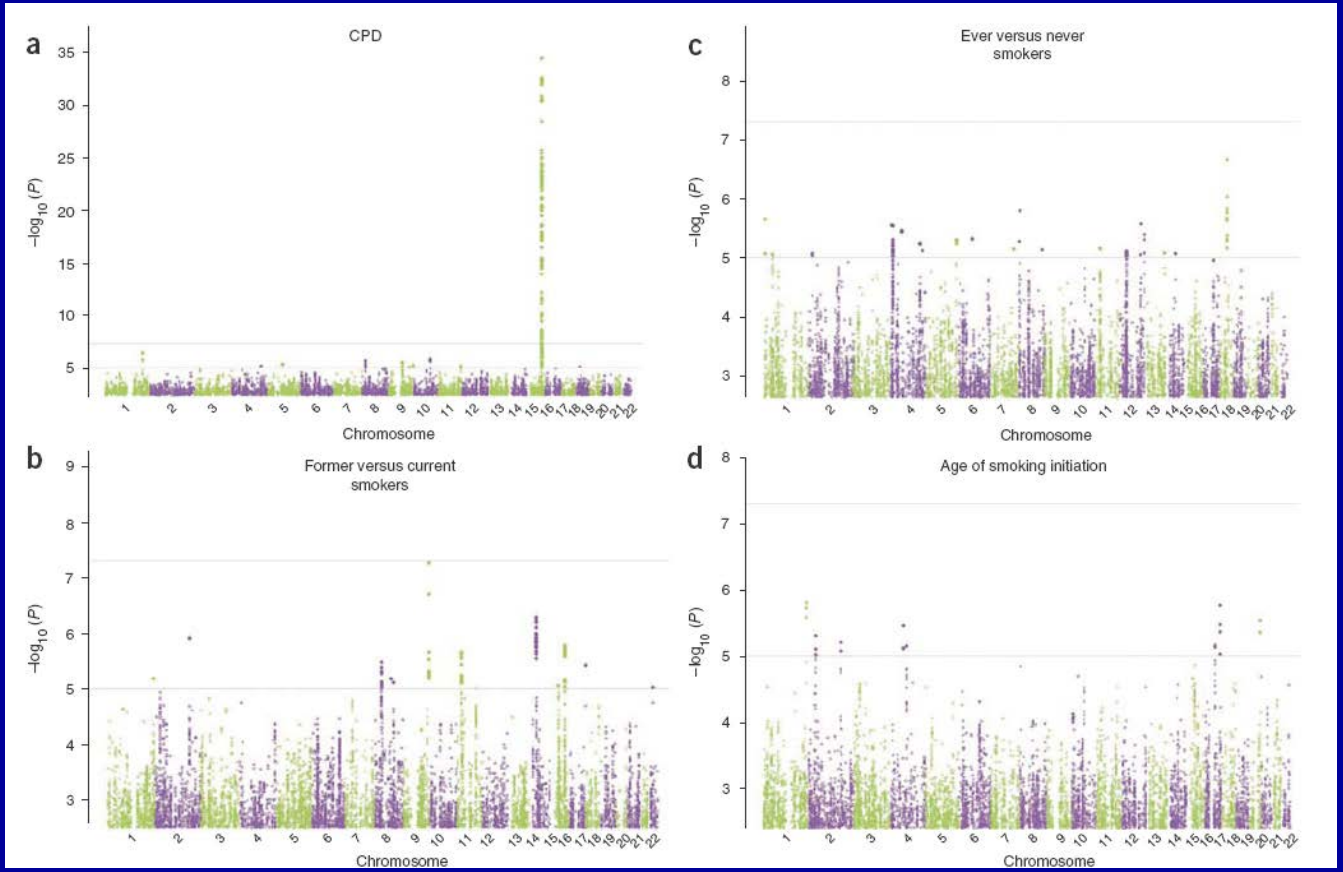
Fig. 2. Bias patterns partitioned by disciplinary domain. Each panel re-

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and other tools)
- Reproducibility practices
- Containment of conflicted sponsors and authors
- More appropriate statistical methods
- Standardization of definitions and analyses
- More stringent thresholds for claiming discoveries or "successes"
- Improvement of study design standards
- Improvements in peer review, reporting, and dissemination of research
- Better training of scientific workforce in methods and statistical literacy

Ioannidis, PLoS Medicine 2014

Large-scale collaboration and adoption of replication culture



Registration

- Level 0: no registration
- Level 1: registration of dataset
- Level 2: registration of protocol
- Level 3: registration of analysis plan
- Level 4: registration of analysis plan and raw data
- Level 5: open live streaming

POLICY FORUM

DATA ACCESS

Toward unrestricted use of public genomic data

Publication interests should not limit access to public data

By **Rudolf I. Amann, Shakuntala Baichoo, Benjamin J. Blencowe, Peer Bork, Mark Borodovsky, Cath Brooksbank, Patrick S. G. Chain, Rita R. Colwell, Daniele G. Daffonchio, Antoine Danchin, Victor de Lorenzo, Pieter C. Dorrestein, Robert D. Finn, Claire M. Fraser, Jack A. Gilbert, Steven J. Hallam, Philip Hugenholtz, John P. A. Ioannidis, Janet K. Jansson, Jihyun F. Kim, Hans-Peter Klenk, Martin G. Klotz, Rob Knight, Konstantinos T. Konstantinidis, Nikos C. Kyrpides, Christopher E. Mason, Alice C. McHardy, Folker Meyer, Christos A. Ouzounis, Aristides A. N. Patrinos, Mircea Podar, Katherine S. Pollard, Jacques Ravel, Alejandro Reyes Muñoz, Richard J. Roberts, Ramon Rosselló-Móra, Susanna-Assunta Sansone, Patrick D. Schloss, Lynn M. Schriml, João C. Setubal, Rotem Sorek, Rick L. Stevens, James M. Tiedje, Adrian Turjanski, Gene W. Tyson, David W. Ussery, George M. Weinstock, Owen White, William B. Whitman, Ioannis Xenarios**

Despite some notable progress in data sharing policies and practices, restrictions are still often placed on the open and unconditional use of various genomic data after they have received official approval for release to the public domain or to public databases. These restrictions, which often conflict with the terms and conditions of the funding bodies who supported the release of those data for the benefit of the scientific community and society, are perpetuated by the lack of clear guiding rules for data usage. Existing guidelines for data released to the public domain recognize but fail to resolve tensions between the importance of free and unconditional use of these data and the "right" of the data producers to the first publication. This self-contradiction has resulted in a loophole that allows different interpretations and a continuous debate between data producers and data users on the use of public data. We argue that the publicly available data should be treated as open data, a shared resource with unrestricted use for analysis, interpretation, and publication.

SHARING, PUBLISHING, PARADOX

The landmark 2003 Fort Lauderdale

platforms (such as genome-wide association studies) or even with wider spectra of data being covered (for example, the 2014 National Institutes of Health Genome Data Sharing Policy). A number of widely adopted developments [such as open-access, FAIR (findable, accessible, interoperable, and reusable) principles (*5*)] have created a more refined data-sharing ecosystem that is not captured by the earlier agreements. In order to address the current complexities of data sharing, new community efforts are being organized. For example, the European Bioinformatics Institute has launched a community survey to determine what most investigators want for open data in microbiome research.

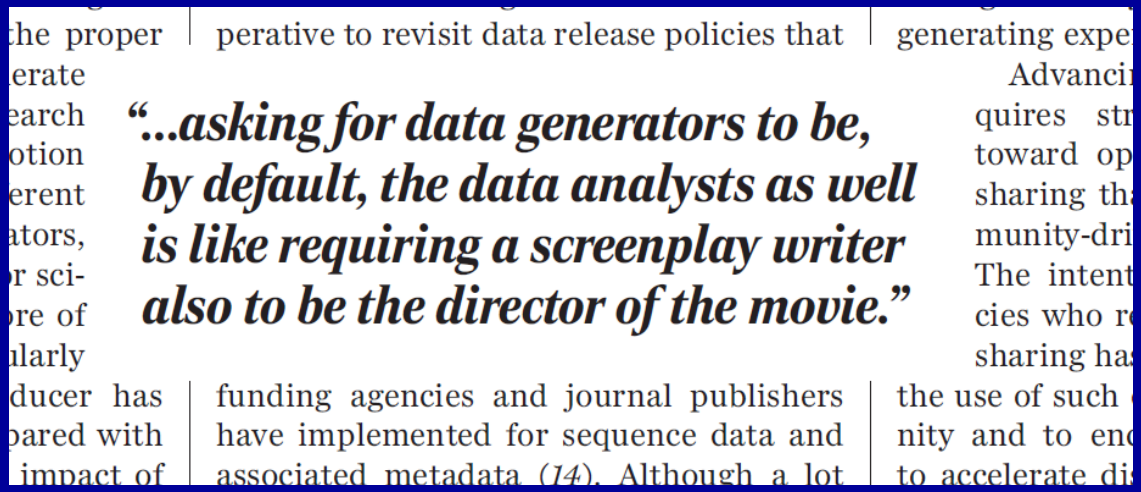
However, despite improvements in aspects of data sharing policies in the past 15 years, with much focus on determining when data should be made publicly available (for example, the ENCODE project has recently eliminated the 9-month moratorium on data usage, applied in earlier phases of the project), policies have not adequately resolved a critical

> ***"Unrestricted use of public data should be aligned with a reward system in research and academia."***

not always guaranteed that they can publish prominent peer-reviewed reports if others use their data first. This paradox is evident despite the agreement's acknowledgment of academic fair play, encouraging users of data publicly released in this fashion to "appropriately cite the source of the data analysed and acknowledge the resource producers."

In light of this, supporters of restricted use of public genomic data point to the agreement to argue that the first use of the data after they become public should still be restricted so that the principal investigator (PI) under whom the data were generated should retain the rights to first publication. This has been frequently implemented as official data release policy from various institutes or research consortia who make the data publicly available but restrict the analysis by the larger community (*6*). Even when the data have become public following the data release policy of the funding agency, proponents of this view argue that outside investigators should still contact the scientist(s) that

Research parasites and movie directors



...he proper ... perative to revisit data release policies that ... generating expe...

...erate ... Advanci...

...earch ... quires str...

...otion ... toward op...

...erent ... sharing tha...

...ators, ... munity-dri...

...r sci- ... The intent...

...re of ... cies who re...

...ularly ... sharing has...

...ducer has ... funding agencies and journal publishers ... the use of such ...

...pared with ... have implemented for sequence data and ... nity and to enc...

...impact of ... associated metadata (*14*). Although a lot ... to accelerate di...

"...asking for data generators to be, by default, the data analysts as well is like requiring a screenplay writer also to be the director of the movie."

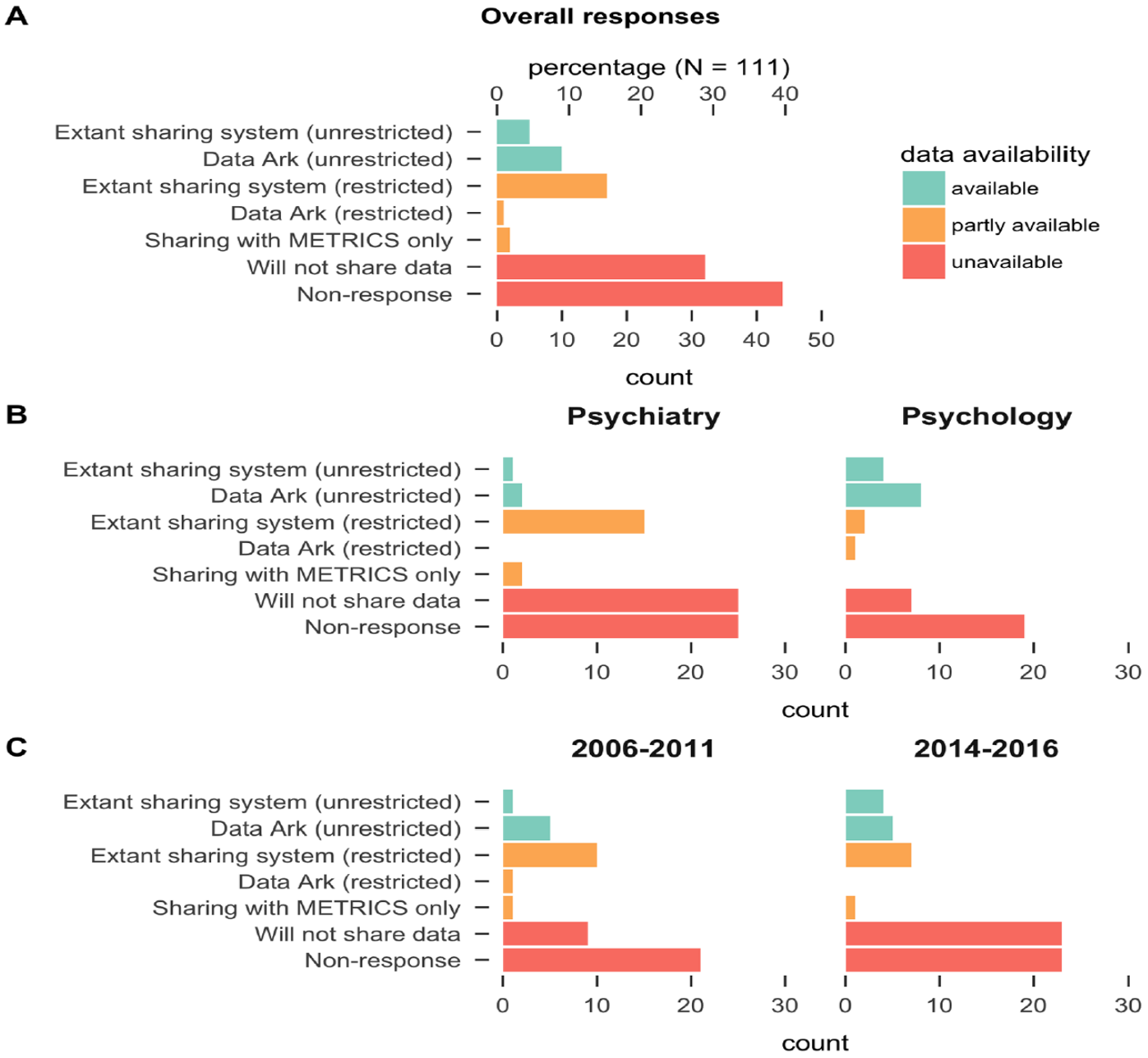


Fig 1. Responses to data request, overall (Panel A) and broken down by field (Panel B) and sample time period (Panel C).

https://doi.org/10.1371/journal.pone.0201856.g001

Hardwicke, Ioannidis 2018

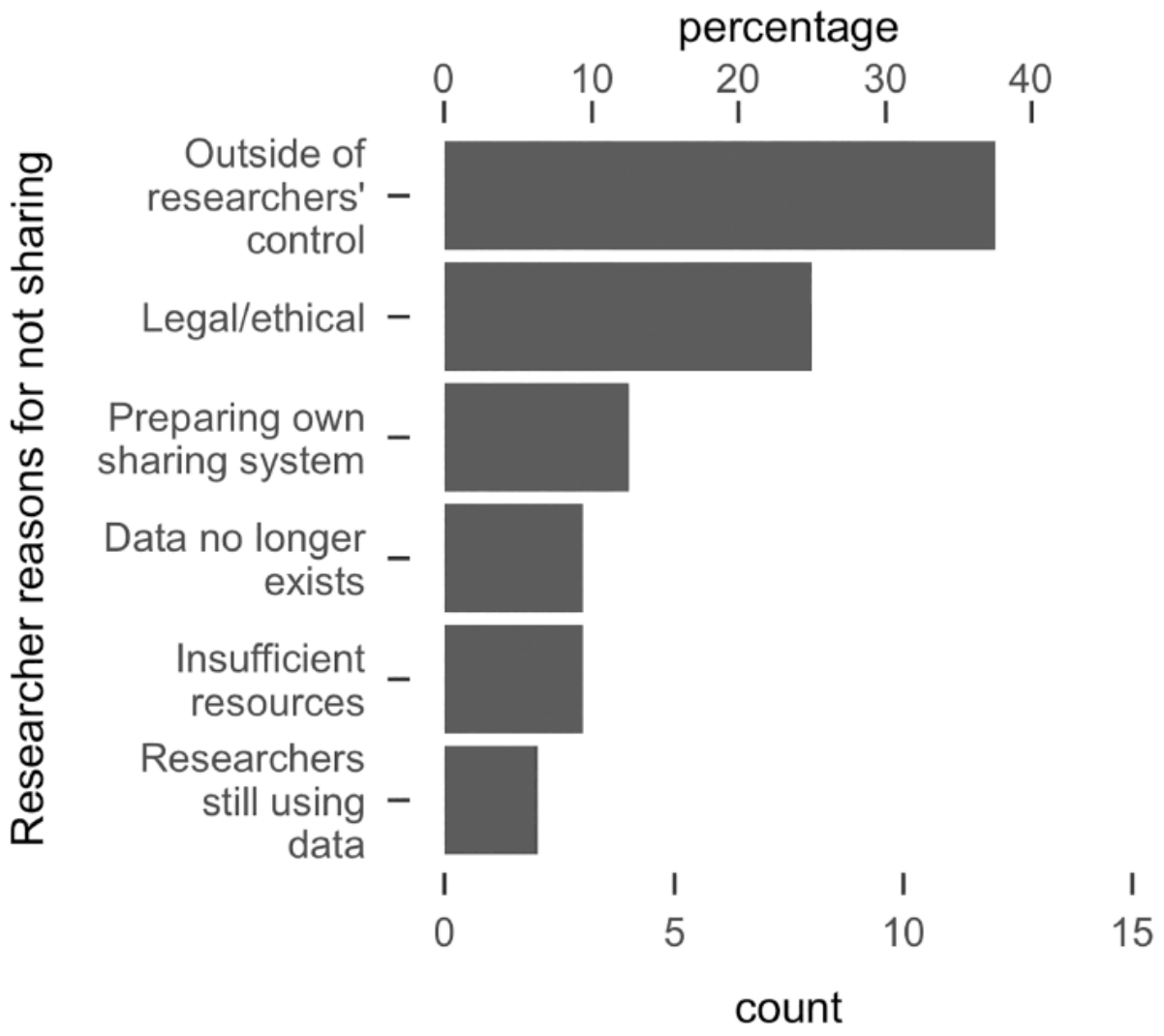


Fig 2. Reasons provided by researchers for not sharing. X-axes represent counts and percentages (of n = 32 who responded that they would not share).

https://doi.org/10.1371/journal.pone.0201856.g002

Transparency: can we trust the data?

Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

Joanna Le Noury,[1] John M Nardo,[2] David Healy,[1] Jon Jureidini,[3] Melissa Raven,[3] Catalin Tufanaru,[4] Elia Abi-Jaoude[5]

ABSTRACT

OBJECTIVES

To reanalyse SmithKline Beecham's Study 329 (published by Keller and colleagues in 2001), the primary objective of which was to compare the efficacy and safety of paroxetine and imipramine with placebo in the treatment of adolescents with unipolar major depression. The reanalysis under the restoring invisible and abandoned trials (RIAT) initiative was done to see whether access to and reanalysis of a full dataset from a randomised controlled trial would have clinically relevant implications for evidence based medicine.

DESIGN

Double blind randomised placebo controlled trial.

SETTING

12 North American academic psychiatry centres, from 20 April 1994 to 15 February 1998.

PARTICIPANTS

275 adolescents with major depression of at least eight weeks in duration. Exclusion criteria included a range of comorbid psychiatric and medical disorders and suicidality.

INTERVENTIONS

Participants were randomised to eight weeks double blind treatment with paroxetine (20-40 mg), imipramine (200-300 mg), or placebo.

MAIN OUTCOME MEASURES

The prespecified primary efficacy variables were change from baseline to the end of the eight week acute treatment phase in total Hamilton depression scale (HAM-D) score and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Prespecified secondary outcomes were changes from baseline to endpoint in depression items in K-SADS-L, clinical global impression, autonomous functioning checklist, self-perception profile, and sickness impact scale; predictors of response; and number of patients who relapse during the maintenance phase. Adverse experiences were to be compared primarily by using descriptive statistics. No coding dictionary was prespecified.

RESULTS

The efficacy of paroxetine and imipramine was not statistically or clinically significantly different from placebo for any prespecified primary or secondary efficacy outcome. HAM-D scores decreased by 10.7 (least squares mean) (95% confidence interval 9.1 to 12.3), 9.0 (7.4 to 10.5), and 9.1 (7.5 to 10.7) points, respectively, for the paroxetine, imipramine and placebo groups (P=0.20). There were clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events in the paroxetine group and cardiovascular problems in the imipramine group.

CONCLUSIONS

Neither paroxetine nor high dose imipramine showed efficacy for major depression in adolescents, and there was an increase in harms with both drugs. Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The reanalysis of Study 329 illustrates the necessity of making primary trial data and protocols available to increase the rigour of the evidence base.

Original Investigation

Reanalyses of Randomized Clinical Trial Data

Shanil Ebrahim, PhD; Zahra N. Sohani, MSc; Luis Montoya, DDS; Arnav Agarwal, BSc; Kristian Thorlund, PhD; Edward J. Mills, PhD; John P. A. Ioannidis, MD, DSc

IMPORTANCE Reanalyses of randomized clinical trial (RCT) data may help the scientific community assess the validity of reported trial results.

OBJECTIVES To identify published reanalyses of RCT data, to characterize methodological and other differences between the original trial and reanalysis, to evaluate the independence of authors performing the reanalyses, and to assess whether the reanalysis changed interpretations from the original article about the types or numbers of patients who should be treated.

DESIGN We completed an electronic search of MEDLINE from inception to March 9, 2014, to identify all published studies that completed a reanalysis of individual patient data from previously published RCTs addressing the same hypothesis as the original RCT. Four data extractors independently screened articles and extracted data.

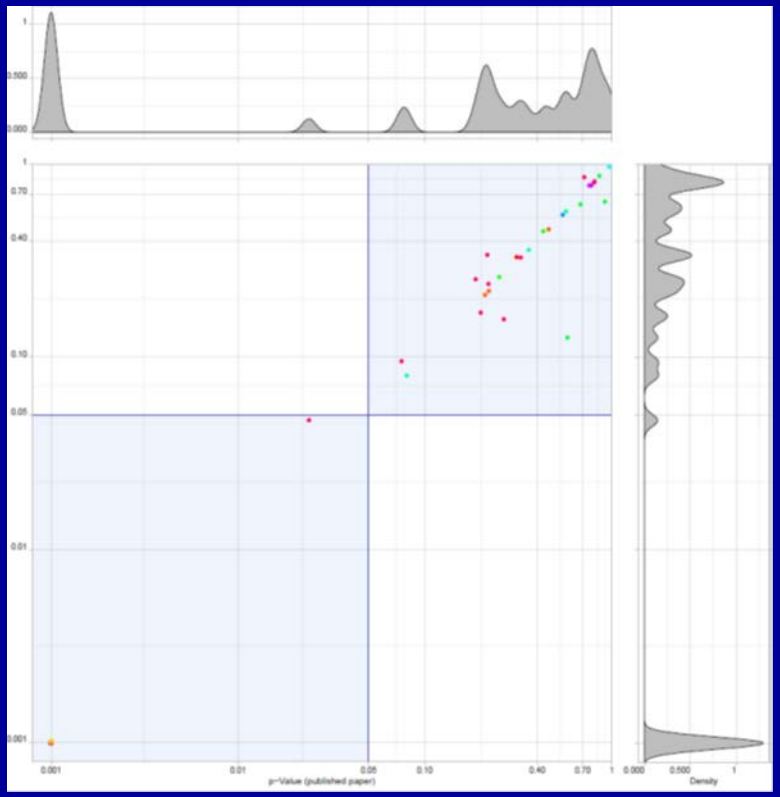
MAIN OUTCOMES AND MEASURES Changes in direction and magnitude of treatment effect, statistical significance, and interpretation about the types or numbers of patients who should be treated.

RESULTS We identified 37 eligible reanalyses in 36 published articles, 5 of which were performed by entirely independent authors (2 based on publicly available data and 2 on data that were provided on request; data availability was unclear for 1). Reanalyses differed most commonly in statistical or analytical approaches (n = 18) and in definitions or measurements of the outcome of interest (n = 12). Four reanalyses changed the direction and 2 changed the magnitude of treatment effect, whereas 4 led to changes in statistical significance of findings. Thirteen reanalyses (35%) led to interpretations different from that of the original article, 3 (8%) showing that different patients should be treated; 1 (3%), that fewer patients should be treated; and 9 (24%), that more patients should be treated.

CONCLUSIONS AND RELEVANCE A small number of reanalyses of RCTs have been published to date. Only a few were conducted by entirely independent authors. Thirty-five percent of published reanalyses led to changes in findings that implied conclusions different from those of the original article about the types and number of patients who should be treated.

JAMA. 2014;312(10):1024-1032. doi:10.1001/jama.2014.9646

46% retrieval rate for raw data of randomized trials under full data



Naudet et al, BMJ 2018

PLOS | BIOLOGY

META-RESEARCH ARTICLE

Reproducible Research Practices and Transparency across the Biomedical Literature

Shareen A. Iqbal[1☯], Joshua D. Wallach[2,3☯], Muin J. Khoury[4,5], Sheri D. Schully[4], John P. A. Ioannidis[2,3,6,7*]

There is a growing movement to encourage reproducibility and transparency practices in the scientific community, including public access to raw data and protocols, the conduct of replication studies, systematic integration of evidence in systematic reviews, and the documentation of funding and potential conflicts of interest. In this survey, we assessed the current status of reproducibility and transparency addressing these indicators in a random sample of 441 biomedical journal articles published in 2000–2014. Only one study provided a full protocol and none made all raw data directly available. Replication studies were rare ($n = 4$), and only 16 studies had their data included in a subsequent systematic review or meta-analysis. The majority of studies did not mention anything about funding or conflicts of interest. The percentage of articles with no statement of conflict decreased substantially between 2000 and 2014 (94.4% in 2000 to 34.6% in 2014); the percentage of articles reporting statements of conflicts (0% in 2000, 15.4% in 2014) or no conflicts (5.6% in 2000, 50.0% in 2014) increased. Articles published in journals in the clinical medicine category versus other fields were almost twice as likely to not include any information on funding and to have private funding. This study provides baseline data to compare future progress in improving these indicators in the scientific literature.

Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017

Joshua D. Wallach[1,2], **Kevin W. Boyack**[3], **John P. A. Ioannidis**[4,5,6,7,8]*

1 Department of Environmental Health Sciences, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Collaboration for Research Integrity and Transparency, Yale School of Medicine, Yale University, New Haven, Connecticut, United States of America, **3** SciTech Strategies, Inc., Albuquerque, New Mexico, United States of America, **4** Stanford Prevention Research Center, Department of Medicine, Stanford University, Stanford, California, United States of America, **5** Department of Health Research and Policy, Stanford University, Stanford, California, United States of America, **6** Department of Biomedical Data Science, Stanford University, Stanford, California, United States of America, **7** Department of Statistics, Stanford University, Stanford, California, United States of America, **8** Meta-Research Innovation Center at Stanford, Stanford University, Stanford, California, United States of America

* jioannid@stanford.edu

Abstract

Currently, there is a growing interest in ensuring the transparency and reproducibility of the published scientific literature. According to a previous evaluation of 441 biomedical journals articles published in 2000–2014, the biomedical literature largely lacked transparency in important dimensions. Here, we surveyed a random sample of 149 biomedical articles published between 2015 and 2017 and determined the proportion reporting sources of public and/or private funding and conflicts of interests, sharing protocols and raw data, and undergoing rigorous independent replication and reproducibility checks. We also investigated what can be learned about reproducibility and transparency indicators from open access data provided on PubMed. The majority of the 149 studies disclosed some information regarding funding (103, 69.1% [95% confidence interval, 61.0% to 76.3%]) or conflicts of interest (97, 65.1% [56.8% to 72.6%]). Among the 104 articles with empirical data in which protocols or data sharing would be pertinent, 19 (18.3% [11.6% to 27.3%]) discussed publicly available data; only one (1.0% [0.1% to 6.0%]) included a link to a full study protocol. Among the 97 articles in which replication in studies with different data would be pertinent, there were five replication efforts (5.2% [1.9% to 12.2%]). Although clinical trial identification numbers and funding details were often provided on PubMed, only two of the articles without a full text article in PubMed Central that discussed publicly available data at the full text level also contained information related to data sharing on PubMed; none had a conflicts of interest statement on PubMed. Our evaluation suggests that although there have been improvements over the last few years in certain key indicators of reproducibility and transparency, opportunities exist to improve reproducible research practices across the biomedical literature and to make features related to reproducibility more readily visible in PubMed.

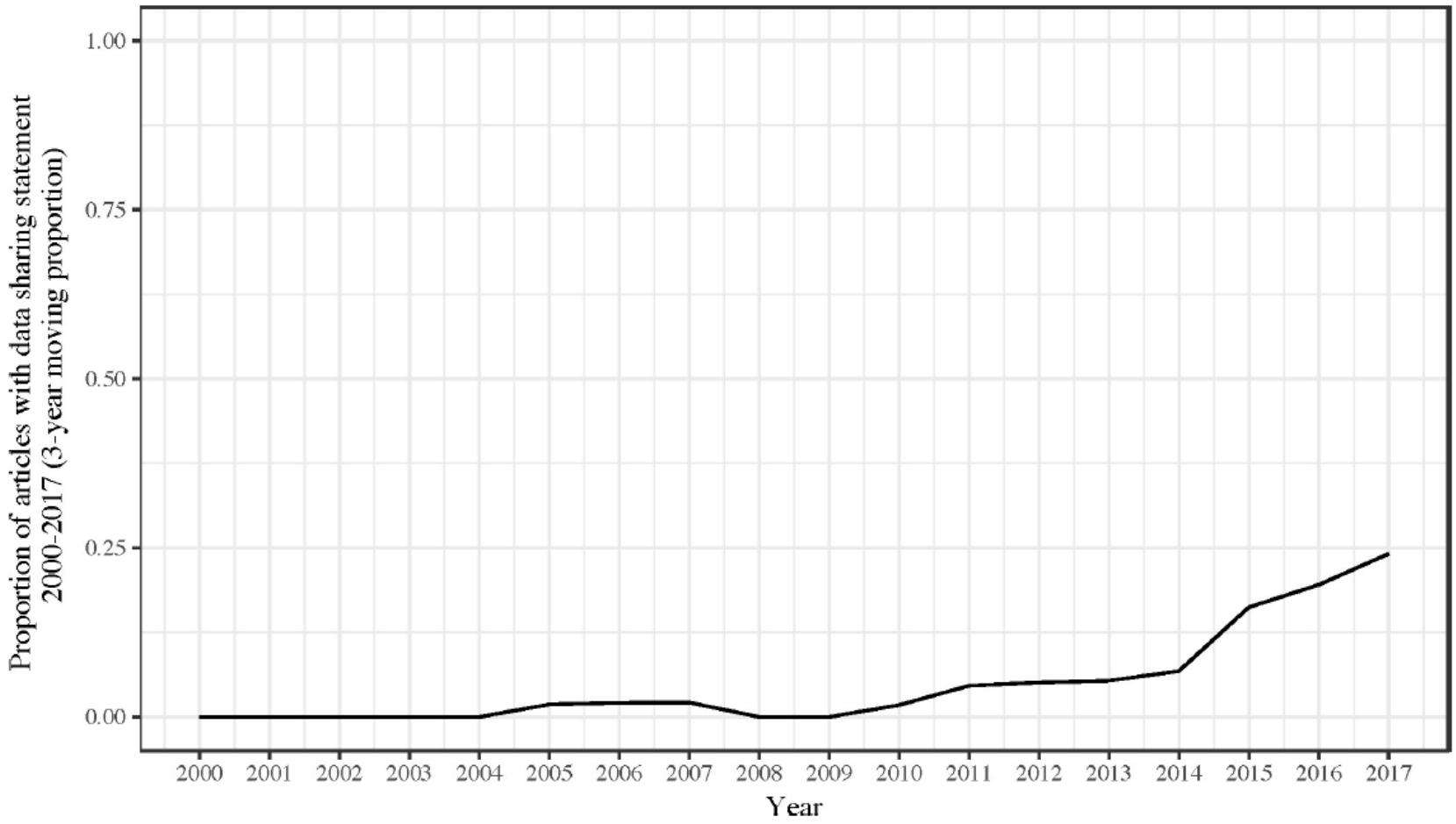


Fig 2. Proportion of articles with data sharing statement, 2000–2017 (3-year moving proportion). Underlying data for Fig 2 can be found at https://osf.io/3ypdn/.

https://doi.org/10.1371/journal.pbio.2006930.g002

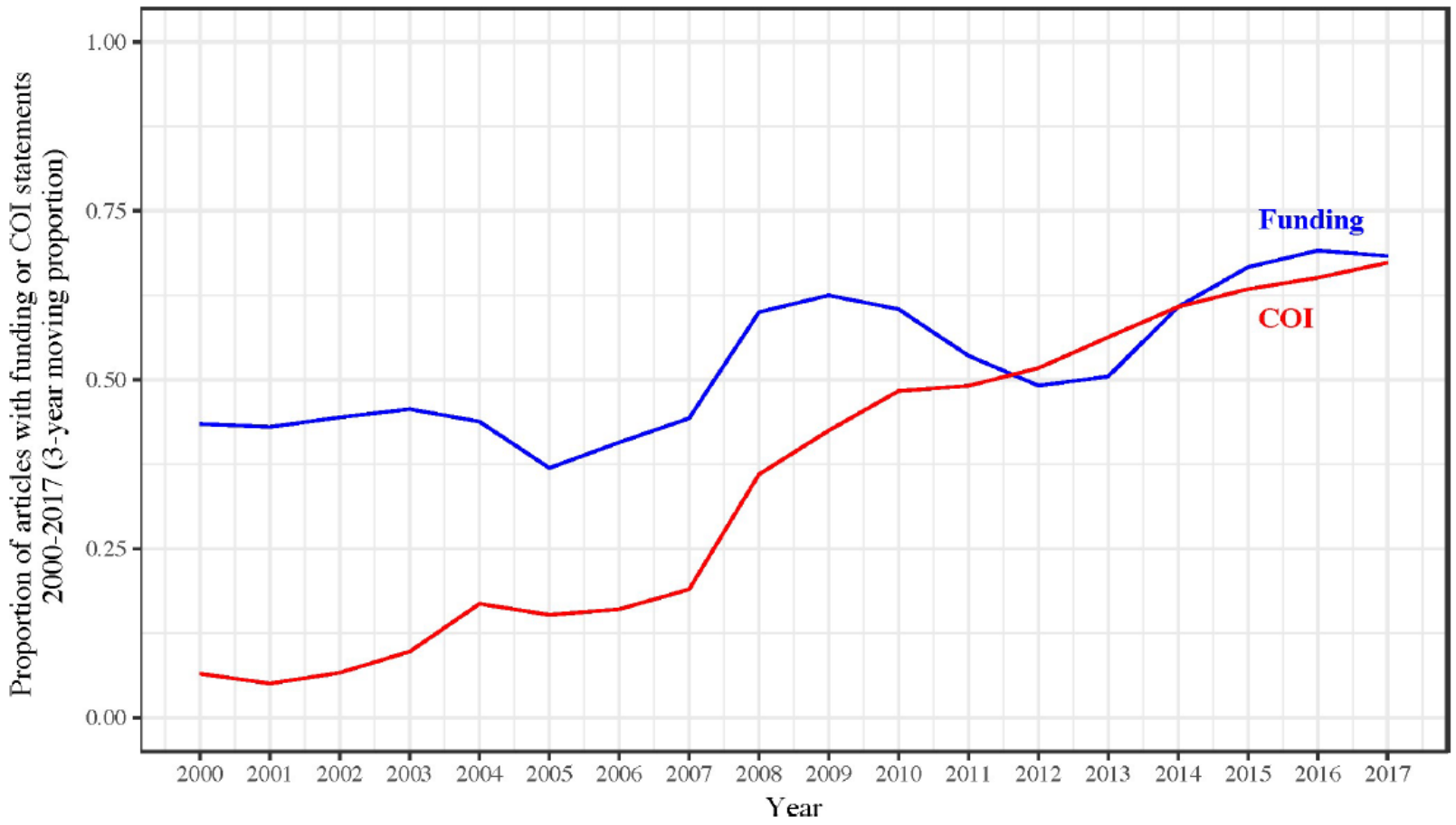


Fig 1. Proportion of articles with funding or COI statements, 2000–2017 (3-year moving proportion). Underlying data for Fig 1 can be found at https://osf.io/3ypdn/. COI, conflicts of interest

https://doi.org/10.1371/journal.pbio.2006930.g001

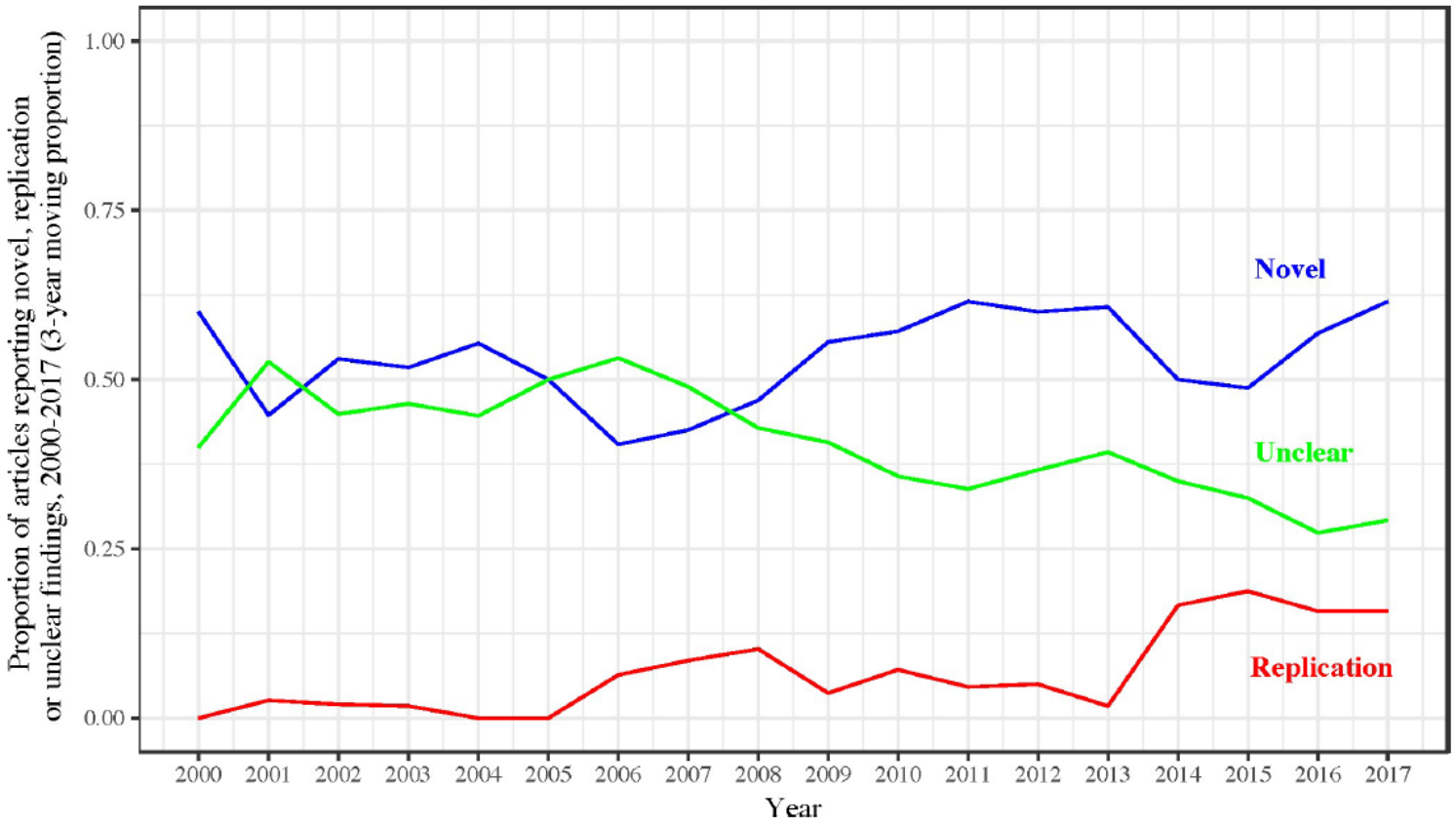
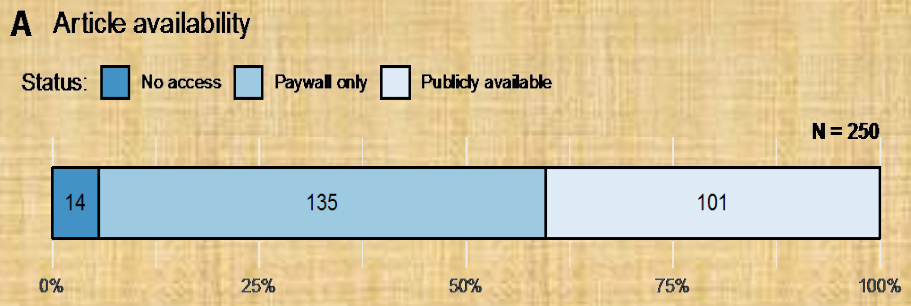
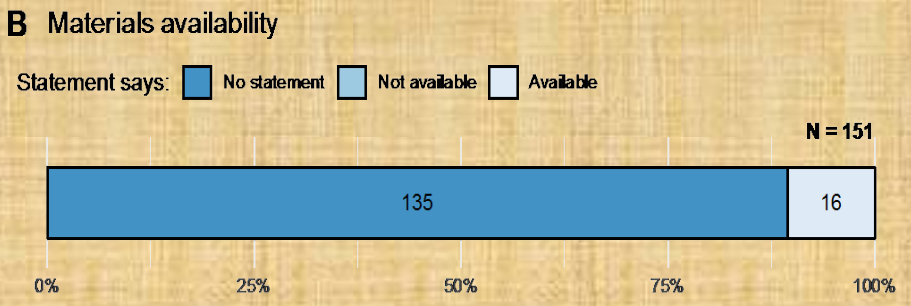


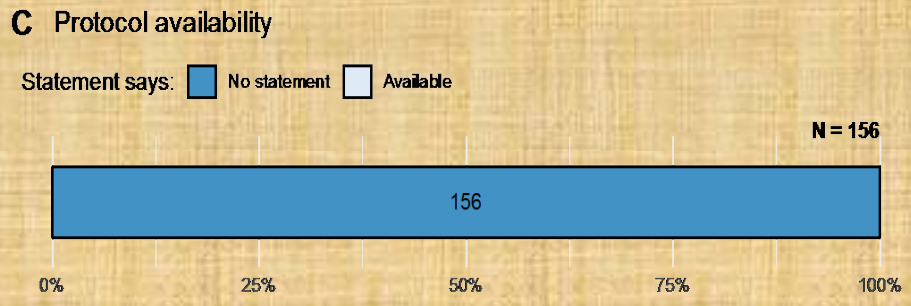
Fig 3. Proportion of articles reporting novel, replication, or unclear findings, 2000–2017 (3-year moving proportion). Underlying data for Fig 3 can be found at https://osf.io/3ypdn/.

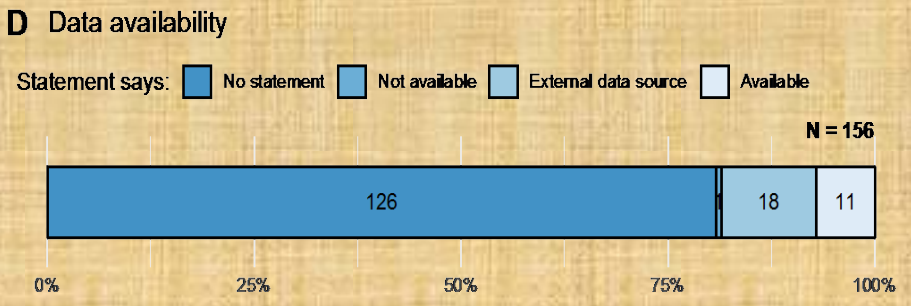
https://doi.org/10.1371/journal.pbio.2006930.g003

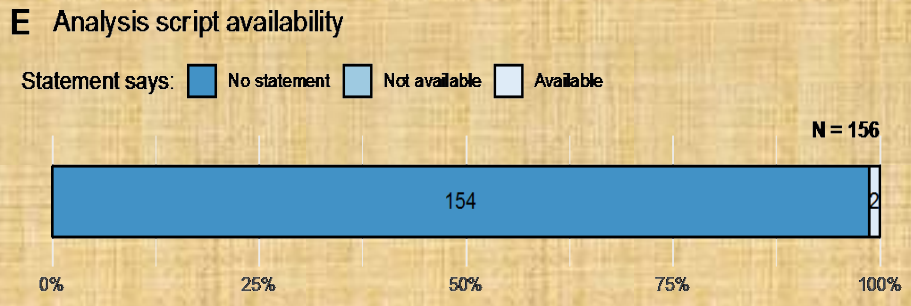
Landscape in social sciences

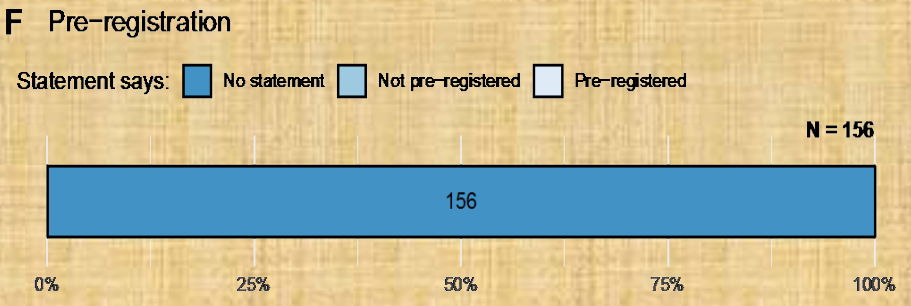


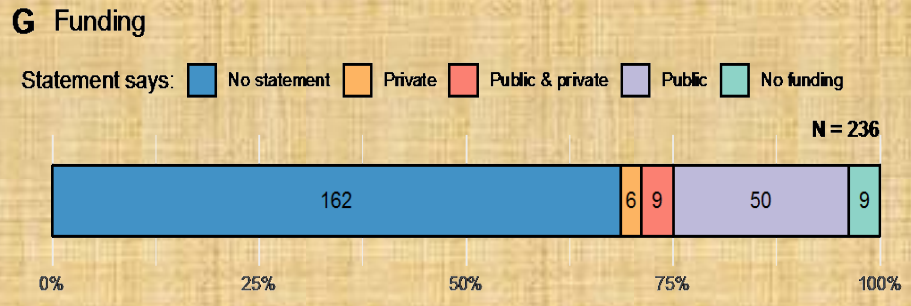


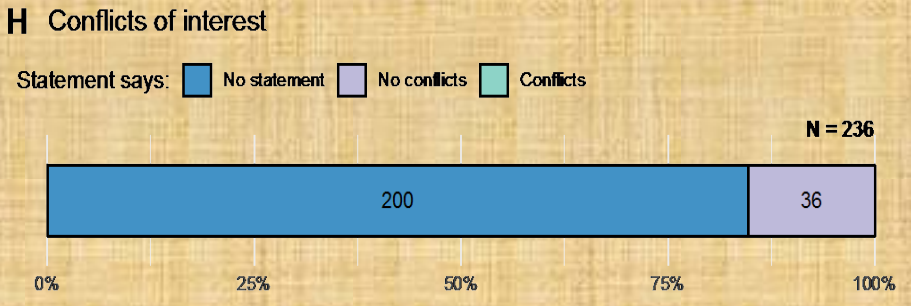












REPRODUCIBILITY

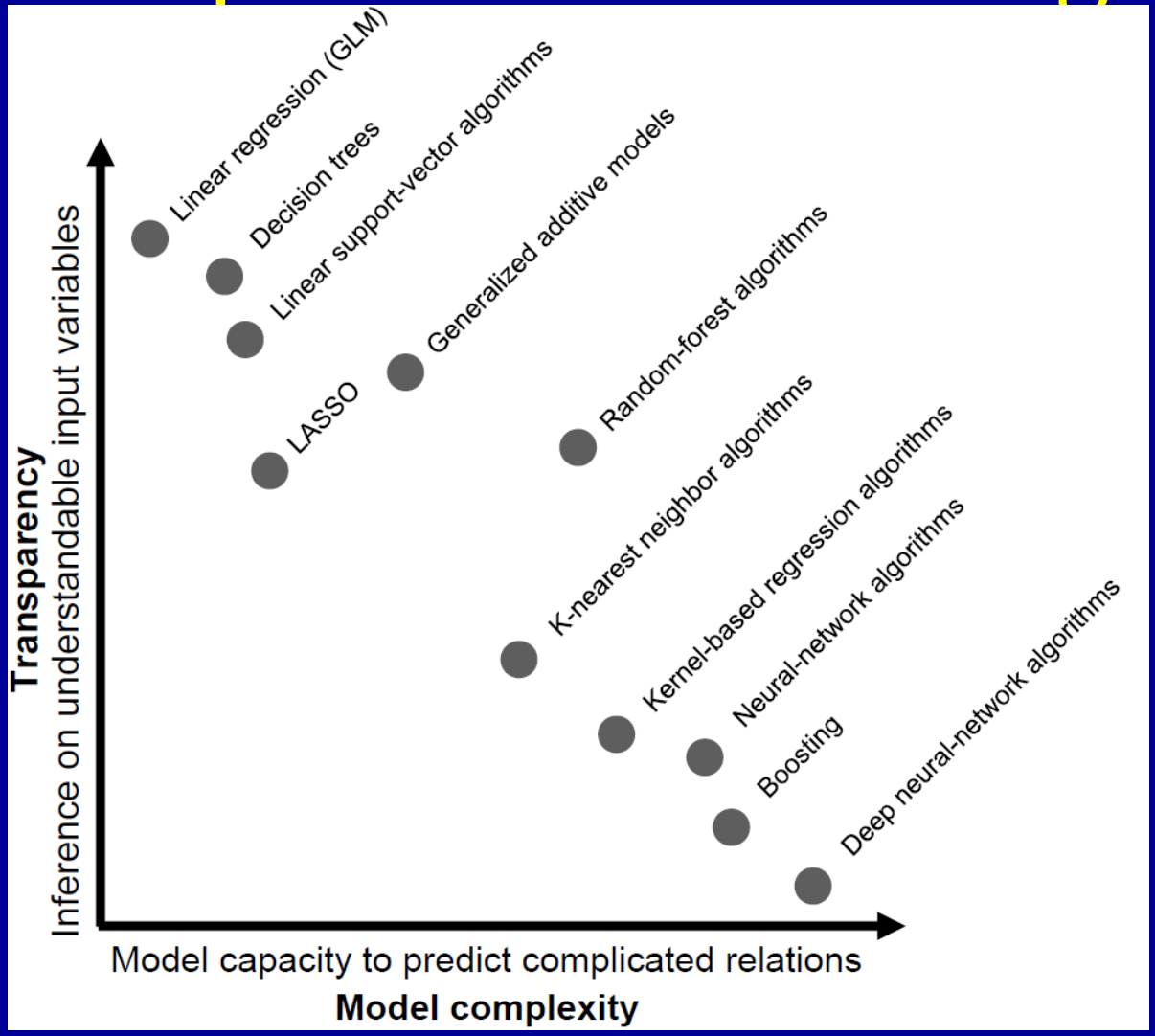
Enhancing Reproducibility for Computational Methods

Data, code and workflows should be available and cited.

By Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer

Science, December 2, 2016

Transparency versus complexity in predictive modeling



Dzok and Ioannidis, Trends in Neuroscience 2019

PLOS | BIOLOGY

COMMUNITY PAGE

Meta-research: Evaluation and Improvement of Research Methods and Practices

John P. A. Ioannidis*, Daniele Fanelli, Debbie Drake Dunne, Steven N. Goodman

Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America

* jioannid@stanford.edu

Table 1. Major themes covered by meta-research.

Meta-research area	Specific interests (nonexhaustive list)
Methods: "performing research"—study design, methods, statistics, research synthesis, collaboration, and ethics	Biases and questionable practices in conducting research, methods to reduce such biases, meta-analysis, research synthesis, integration of evidence, crossdesign synthesis, collaborative team science and consortia, research integrity and ethics
Reporting: "communicating research"—reporting standards, study registration, disclosing conflicts of interest, information to patients, public, and policy-makers	Biases and questionable practices in reporting, explaining, disseminating and popularizing research, conflicts of interest disclosure and management, study registration and other bias-prevention measures, and methods to monitor and reduce such issues
Reproducibility: "verifying research"—sharing data and methods, repeatability, replicability, reproducibility, and self-correction	Obstacles to sharing data and methods, replication studies, replicability and reproducibility of published research, methods to improve them, effectiveness of correction and self-correction of the literature, and methods to improve them
Evaluation: "evaluating research"—prepublication peer review, postpublication peer review, research funding criteria, and other means of evaluating scientific quality	Effectiveness, costs, and benefits of old and new approaches to peer review and other science assessment methods, and methods to improve them
Incentives: "rewarding research": promotion criteria, rewards, and penalties in research evaluation for individuals, teams, and institutions	Accuracy, effectiveness, costs, and benefits of old and new approaches to ranking and evaluating the performance, quality, value of research, individuals, teams, and institutions

doi:10.1371/journal.pbio.1002264.t001

Ioannidis, Fanelli, Dunne, Goodman, PLoS Biology 2015

Modeling a (mal)functional universe of science

PLOS BIOLOGY

PERSPECTIVE

The credibility crisis in research: Can economics tools help?

Thomas Gall[1], **John P. A. Ioannidis**[2], **Zacharias Maniadis**[1]*

1 Economics Department, School of Social Sciences, University of Southampton, Southampton, United Kingdom, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America

* z.maniadis@soton.ac.uk

Abstract

The issue of nonreplicable evidence has attracted considerable attention across biomedical and other sciences. This concern is accompanied by an increasing interest in reforming research incentives and practices. How to optimally perform these reforms is a scientific problem in itself, and economics has several scientific methods that can help evaluate research reforms. Here, we review these methods and show their potential. Prominent among them are mathematical modeling and laboratory experiments that constitute affordable ways to approximate the effects of policies with wide-ranging implications.

Check for updates

$$\begin{pmatrix} S_{D+} \\ S_{C+} \\ S_{U+} \end{pmatrix} = D_R \begin{pmatrix} p_T + \epsilon p_F \\ p_T + \alpha\epsilon p_F \\ p_T + \epsilon p_F + \delta \end{pmatrix}$$

$$\begin{pmatrix} S_{D-} \\ S_{C-} \\ S_{U-} \end{pmatrix} = D_R \mathrm{n} \begin{pmatrix} \beta_D \\ \beta_C \\ \beta_U \end{pmatrix}.$$

$$\begin{pmatrix} \nu_P(t) \\ \nu_N(t) \end{pmatrix} = \begin{pmatrix} \frac{JB}{x(t)S_{D+} + y(t)S_{C+} + z(t)S_{U+}} \\ \frac{J(1-B)}{x(t)S_{D-} + y(t)S_{C-} + z(t)S_{U-}} \end{pmatrix}.$$

$$\begin{pmatrix} L_D(t) \\ L_C(t) \\ L_U(t) \end{pmatrix} = \nu_P(t) \begin{pmatrix} S_{D+} \\ S_{C+} \\ S_{U+} \end{pmatrix} + \nu_N(t) \begin{pmatrix} S_{D-} \\ S_{C-} \\ S_{U-} \end{pmatrix}.$$

$$A(t) = \frac{J}{x(t) + y(t) + z(t)}.$$

$$\begin{pmatrix} x(t+1) \\ y(t+1) \\ z(t+1) \end{pmatrix} = \begin{pmatrix} \frac{L_D(t)}{A(t)} x(t) \\ \frac{L_C(t)}{A(t)} y(t) \\ \frac{L_U(t)}{A(t)} z(t) \end{pmatrix}.$$

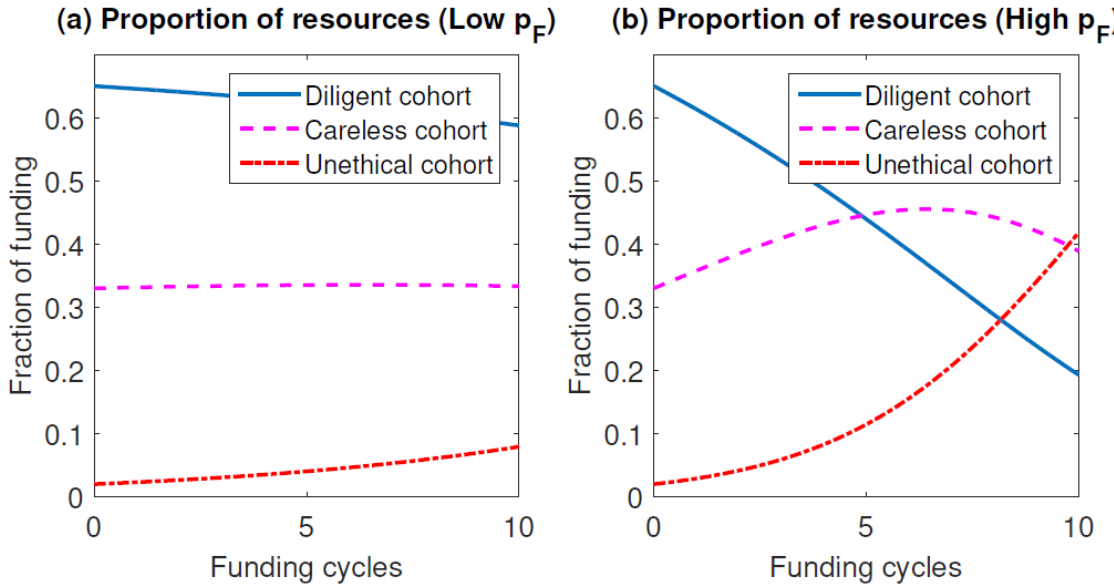
$$\begin{pmatrix} x(t+1) \\ y(t+1) \\ z(t+1) \end{pmatrix} = \begin{pmatrix} \frac{L_D(t)}{A(t)} x(t) + f_D G \\ \frac{L_C(t)}{A(t)} y(t) + f_C G \\ \frac{L_U(t)}{A(t)} z(t) + f_U G \end{pmatrix}.$$

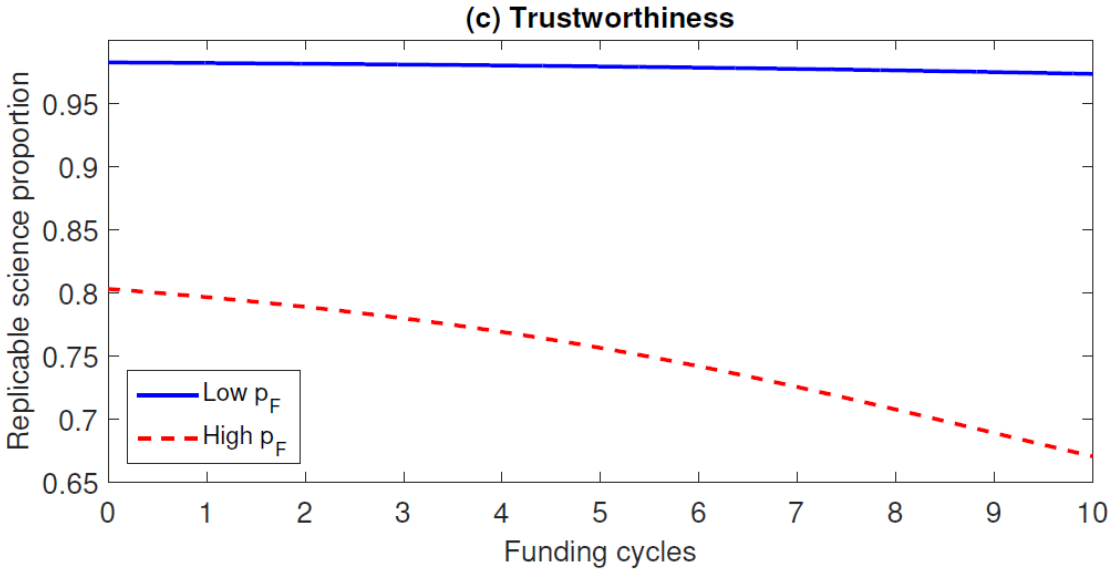
$$z(t+1) = \left(\frac{L_U(t)}{A(t)} - D_R \eta \delta \nu_p(t) \right) z(t) + f_U G$$

$$x(t+1) = \frac{L_D(t)}{A(t)} x(t) + f_D G + R_W$$

$$\nu(t) = \frac{J}{x(t)(S_{D+} + S_{D-}) + y(t)(S_{C+} + S_{C-}) + z(t)(S_{U+} + S_{U-})}$$

$$T(t) = 1 - \frac{\nu_p D_R (x(\epsilon p_F) + y(\alpha\epsilon p_F) + z(\epsilon p_F + \delta))}{J}$$





Grimes, Bauch, Ioannidis. Royal Society Open Science, 2018

Re-engineering the reward system

Table. PQRST Index for Appraising and Rewarding Research

Item in PQRST Index	Operationalization	
	Example	Data Source
P (productivity)	Number of publications in the top tier % of citations for the scientific field and year	ISI Essential Science Indicators (automated)
	Proportion of funded proposals that have resulted in ≥1 published reports of the main results	Funding agency records and automated recording of acknowledged grants (eg, PubMed)
	Proportion of registered protocols that have been published 2 y after the completion of the studies;	Study registries such as ClinicalTrials.gov for trials
Q (quality of scientific work)	Proportion of publications that fulfill ≥1 quality standards	Need to select standards (different per field/design) and may then automate to some extent; may limit to top-cited articles, if cumbersome
R (reproducibility of scientific work)	Proportion of publications that are reproducible	No wide-coverage automated database currently, but may be easy to build, especially if limited to the top-cited pivotal papers in each field.
S (sharing of data and other resources)	Proportion of publications that share their data, materials, and/or protocols (whichever items are relevant)	No wide-coverage automated database currently, but may be easy to build, eg, embed in PubMed at the time of creation of PubMed record and update if more is shared later
T (translational impact of research)	Proportion of publications that have resulted in successful accomplishment of a distal translational milestone, eg, getting promising results in human trials for intervention tested in animals or cell cultures, or licensing of intervention for clinical trials	No wide-coverage automated database currently, would need to be curated by appraiser (eg, funding agency) and may need to be limited to top-cited papers, if cumbersome

Ioannidis and Khoury, JAMA 2014

A user's guide to inflated and manipulated impact factors

John P. A. Ioannidis[1,2,3,4] | Brett D. Thombs[5,6]

TABLE 2 Key measures that capture mechanisms of JIF inflation[a]

Journal	Self-citing boost	Skewness & nonarticle inflation	Expert-based blockbusters
Nature	1	66	0
Science	1	96	0
PLoS Medicine	3	95	0
New England Journal of Medicine	1	120	0
JAMA	2	107	1
British Medical Journal	6	237	0
Journal of Clinical Epidemiology	12	112	0
EJCI	7	54	0
European Heart Journal	7	134	6
Revista Espanola de Cardiologia	52	417	0
European Journal of Heart Failure	20	78	1
Europace	15	162	2

Assessing scientists for hiring, promotion, and tenure

David Moher[1,2]*, Florian Naudet[2,3], Ioana A. Cristea[2,4], Frank Miedema[5], John P. A. Ioannidis[2,6,7,8,9], Steven N. Goodman[2,6,7]

1 Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America, **3** INSERM CIC-P 1414, Clinical Investigation Center, CHU Rennes, Rennes 1 University, Rennes, France, **4** Department of Clinical Psychology and Psychotherapy, Babeş-Bolyai University, Cluj-Napoca, Romania, **5** Executive Board, UMC Utrecht, Utrecht University, Utrecht, the Netherlands, **6** Department of Medicine, Stanford University, Stanford, California, United States of America, **7** Department of Health Research and Policy, Stanford University, Stanford, California, United States of America, **8** Department of Biomedical Data Science, Stanford University, Stanford, California, United States of America, **9** Department of Statistics, Stanford University, Stanford, California, United States of America

* dmoher@ohri.ca

Abstract

Assessment of researchers is necessary for decisions of hiring, promotion, and tenure. A burgeoning number of scientific leaders believe the current system of faculty incentives and rewards is misaligned with the needs of society and disconnected from the evidence about the causes of the reproducibility crisis and suboptimal quality of the scientific publication record. To address this issue, particularly for the clinical and life sciences, we convened a 22-member expert panel workshop in Washington, DC, in January 2017. Twenty-two academic leaders, funders, and scientists participated in the meeting. As background for the meeting, we completed a selective literature review of 22 key documents critiquing the current incentive system. From each document, we extracted how the authors perceived the problems of assessing science and scientists, the unintended consequences of maintaining the status quo for assessing scientists, and details of their proposed solutions. The resulting table was used as a seed for participant discussion. This resulted in six principles for assessing scientists and associated research and policy implications. We hope the content of this paper will serve as a basis for establishing best practices and redesigning the current approaches to assessing scientists by the many players involved in that process.

PLOS BIOLOGY

COMMUNITY PAGE

A standardized citation metrics author database annotated for scientific field

John P. A. Ioannidis[1]*, Jeroen Baas[2], Richard Klavans[3], Kevin W. Boyack[4]

Table 1. Percentiles of total citations and composite citation metric for each of 22 large scientific fields, career-long data (citations from 1996–2017). Total citations include self-citations.

Scientific field	Authors	Percentile, total citations				Percentile, composite index			
		25th	50th	75th	90th	25th	50th	75th	90th
Agriculture, Fisheries, & Forestry	232,801	32	90	255	671	0.997	1.418	1.892	2.394
Built Environment & Design	36,534	17	51	143	370	0.953	1.344	1.821	2.335
Enabling & Strategic Technologies	475,142	23	75	233	678	0.890	1.330	1.807	2.300
Engineering	436,723	18	56	174	499	0.896	1.316	1.794	2.314
Information & Communication Technologies	339,284	20	60	193	574	0.970	1.380	1.862	2.383
Communication & Textual Studies	20,292	12	32	91	240	1.141	1.542	1.995	2.430
Historical Studies	25,277	16	40	105	263	1.138	1.568	2.012	2.429
Philosophy & Theology	13,861	12	32	87	217	1.145	1.558	2.003	2.453
Visual & Performing Arts	3,717	7	17	40	83	0.985	1.316	1.680	1.998
Economics & Business	108,277	28	83	258	708	1.191	1.651	2.194	2.730
Social Sciences	119,260	20	56	158	423	1.159	1.606	2.114	2.615
General Science & Technology	69,789	14	41	122	399	0.735	1.030	1.392	1.760
General Arts, Humanities, & Social Sciences	4,091	11	28	70	158	1.026	1.403	1.810	2.192
Biomedical Research	626,753	68	212	641	1,769	1.095	1.598	2.111	2.660
Clinical Medicine	2,113,734	41	141	467	1,430	0.935	1.420	1.979	2.568
Psychology & Cognitive Sciences	96,159	41	128	403	1,198	1.189	1.641	2.198	2.842
Public Health & Health Services	141,162	31	92	273	785	0.988	1.427	1.949	2.520
Biology	236,108	47	140	426	1,178	1.151	1.603	2.125	2.686
Chemistry	506,526	45	129	362	989	1.057	1.503	1.967	2.467
Earth & Environmental Sciences	223,246	40	126	405	1,192	1.096	1.562	2.120	2.709
Mathematics & Statistics	96,619	18	52	162	457	1.049	1.503	2.059	2.596
Physics & Astronomy	667,255	38	128	480	1,741	1.022	1.495	2.042	2.615
Unassigned*	287,779	2	7	18	42	0.463	0.672	0.985	1.302
TOTAL	6,880,389	29	102	346	1,077	0.946	1.420	1.951	2.513

A manifesto for reproducible science

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

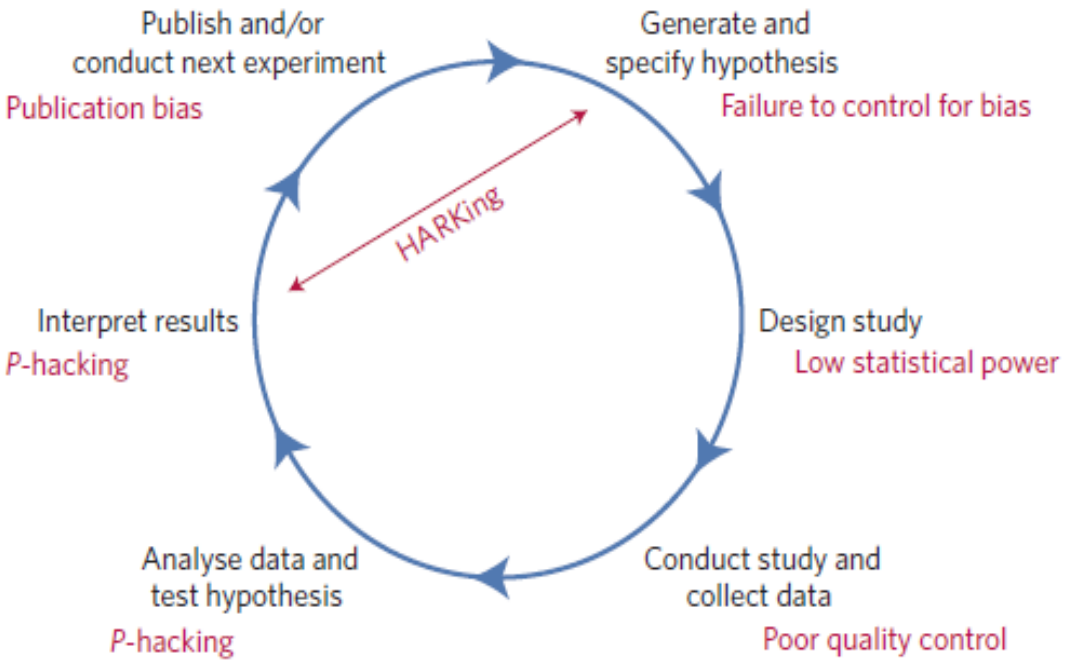


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various

Understand and align interests of stakeholders

Table 1. Some major stakeholders in science and their extent of interest in research and its results from various perspectives; typical patterns are presented (exceptions do occur).

	Extent of interest in research results			
	Publishable	**Fundable**	**Translatable**	**Profitable**
Scientists	+++	+++	+	
Industry – sales and marketing				+++
Industry – R & D			+++	+++
Private investors, including hedge funds			++	+++
Public funders – open (e.g. NIH, NSF)	++		+	
Public funders – closed (e.g. military)			+++	
Not-for-profit funders/philanthropists	++		+++	
Journal editors	+++			+
For-profit publishers	+			+++
Professional and scientific societies	+			
Universities	+	+++		+
Not-for-profit research institutions	+++	+++	+	+
Supporting non-scientific staff		+++		
Hospitals and other professional facilities offering services related to science			+	+++
Other financial entities that are affected by these services (e.g. insurance)				+++
Governments and state/federal authorities				++
Consumers of products and services			+++	

doi:10.1371/journal.pmed.1001747.t001

Concluding comments

- Reproducibility is a central hallmark of research quality and of its potential to translate to useful applications
- The reproducibility of many disciplines of scientific investigation has substantial room for improvement.
- There are many possible interventions that may improve the efficiency of research practices and the reproducibility of the evidence.
- Transparency, openness and sharing are likely to help, but details on “how to” can be important.
- The landscape of reproducibility is currently changing and may change more markedly in the next few years

Special thanks

Uli Dirnagl
Steve Goodman
Shanil Ebrahim
Despina Contopoulos-Ioannidis
Georgia Salanti
Chirag Patel
Lars Hemkens
Ann Hsing
Lamberto Manzoli
Maria Elena Flacco
George Siontis
Denes Szucs
Kostas Siontis
Vangelis Evangelou
Kristin Sainani
Muin Khoury
Orestis Panagiotou
Florence Bourgeois

Special thanks

Joseph Lau
Malcolm MacLeod
Marcus Munafo
David Allison
Josh Wallach
Fotini Karassa
Athina Tatsioni
Evi Ntzani
Ioanna Tzoulaki
Demos Katritsis
Nikos Patsopoulos
Fainia Kavvoura
Brian Nosek
Victoria Stodden
Ele Zeggini
Belinda Burford
Kostas Tsilidis
Jodi Prochaska

Special thanks

Charitini Stavropoulou
Evropi Theodoratou
Nikos Pandis
Huseyin Naci
Vanesa Bellou
Antony Doufas
Lazaros Belbasis
Chris Doucouliagos
Stelios Serghiou
Anna Chaimani
Fotini Chatzinasiou
Stephania Papatheodorou
Florian Naudet
Tom Hardwicke
Perrine Janiaud
Ioana-Alina Cristea
Shannon Brownlee
Vikas Saini
Matthias Egger
Patrick Bossuyt
Andre Uitterlinden
Doug Altman
Deb Zarin
Katherine Flegal

Special thanks

Shanthi Kapaggoda
Ewoud Schuit
Stefania Boccia
David Chavalarias
Jennifer Ware
Viswam Nair
Stephan Bruns
Dorothy Bishop
Tom Trikalinos
Kristina Sundquist
Johanna Int'hout
Kevin Boyack
Brett Thombs
Raj Manrai
Nazmus Saquib
Elizabeth Iorns
Abraham Verghese
Euan Ashley