

# Reproducible research: progress made across disciplines

**Steneck-Mayer lecture, 8<sup>th</sup> WCRI, Athens, June 2024**

4

John P.A. Ioannidis, MD, DSc

Professor of Medicine, of Epidemiology and Population Health, and (by  
courtesy) of Biomedical Data Science, and of Statistics

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

# Disclosures

- My main conflict of interest is that I am studying biases therefore it is very likely that I am biased.
- Don't take what I say for granted!

# Some pre-emptive comments

- Science is the best thing that can happen to humans.
- Most scientific research done to-date has used non-reproducible, suboptimal research practices.
- Science is becoming more massive and more complex. Scientific publications (200 million already and increasing at a rate of >7 million per year) are mostly advertisements (“trust me, this research was done”). Raw data and experimental materials and algorithms usually are not shared.
- Reward systems in academia and science in general are aligned with non-reproducible, suboptimal research practices
- At the same time, the last decade has seen a flurry of interest and multiple efforts to address problems of reproducibility, and improve research practices
- Are we doing better? Can we do better?

Maps of science suggest there are many thousands  
of scientific disciplines.

Their research practices vary substantially

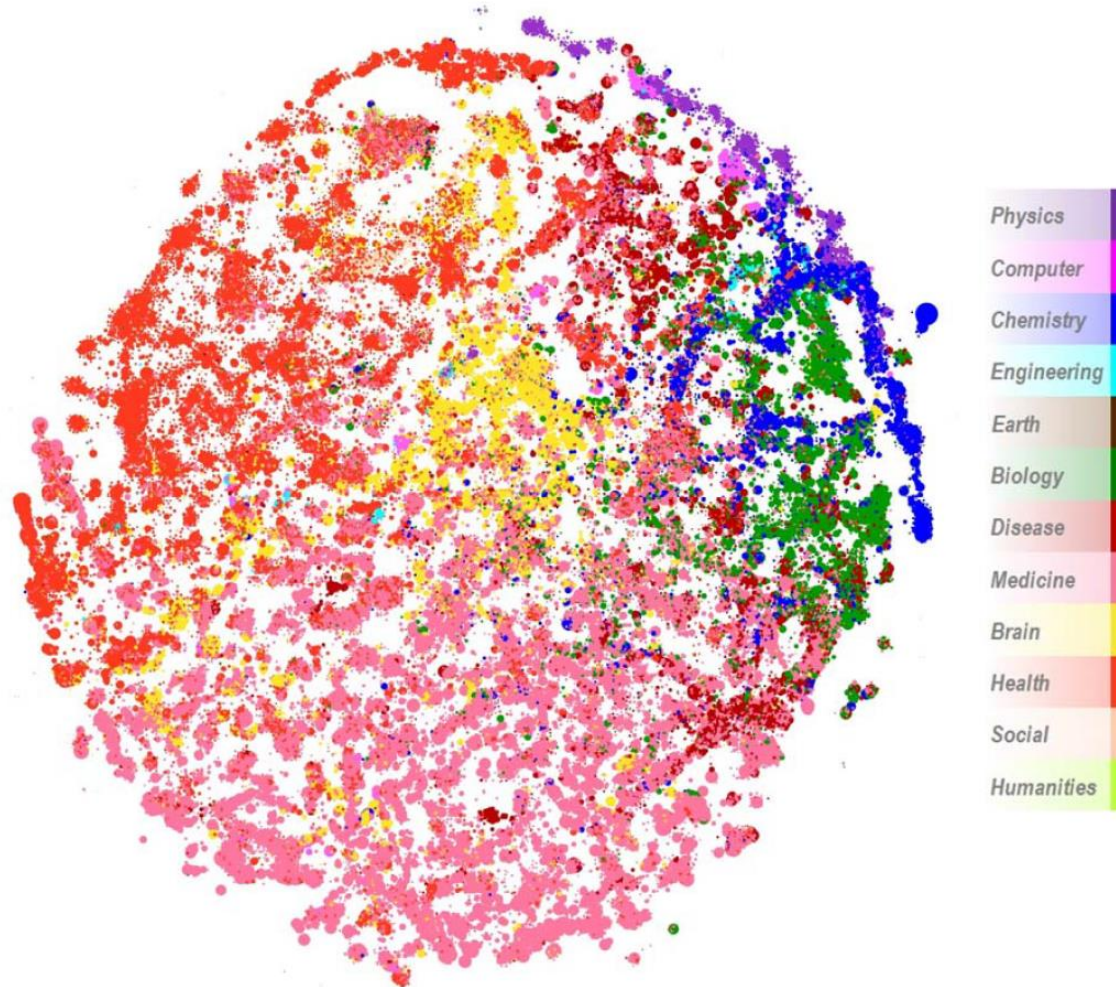


Figure 8. Visual map of the topics in the 12PM<sup>5</sup> model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic.

The published literature is only part of this universe. Much (most?) of this universe is unpublished “dark” matter

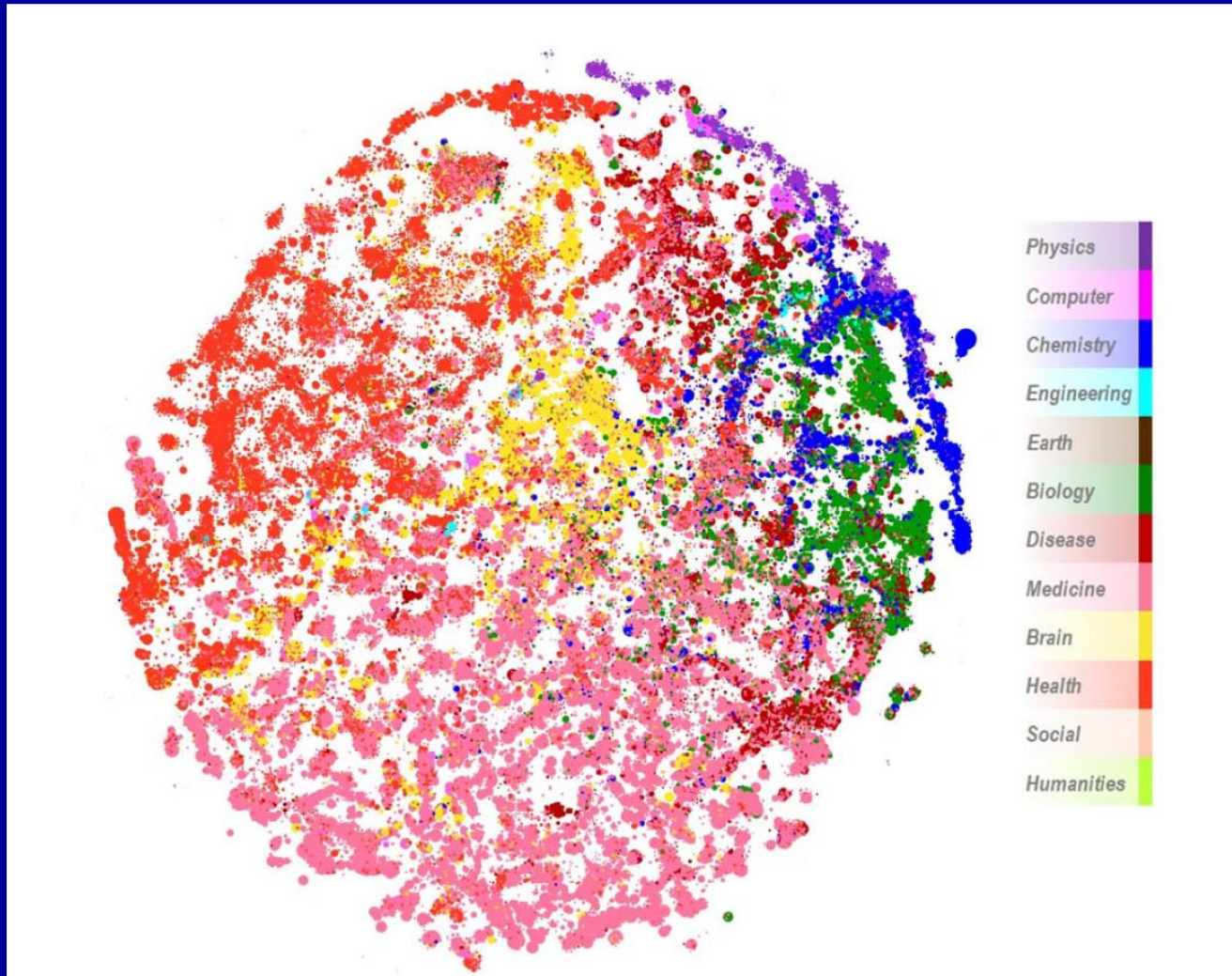


Figure 8. Visual map of the topics in the 12PM<sup>5</sup> model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic.

Even if single fields improve over time, the total universe may get worse if dark matter increases and/or if the least reproducible fields grow at a faster pace

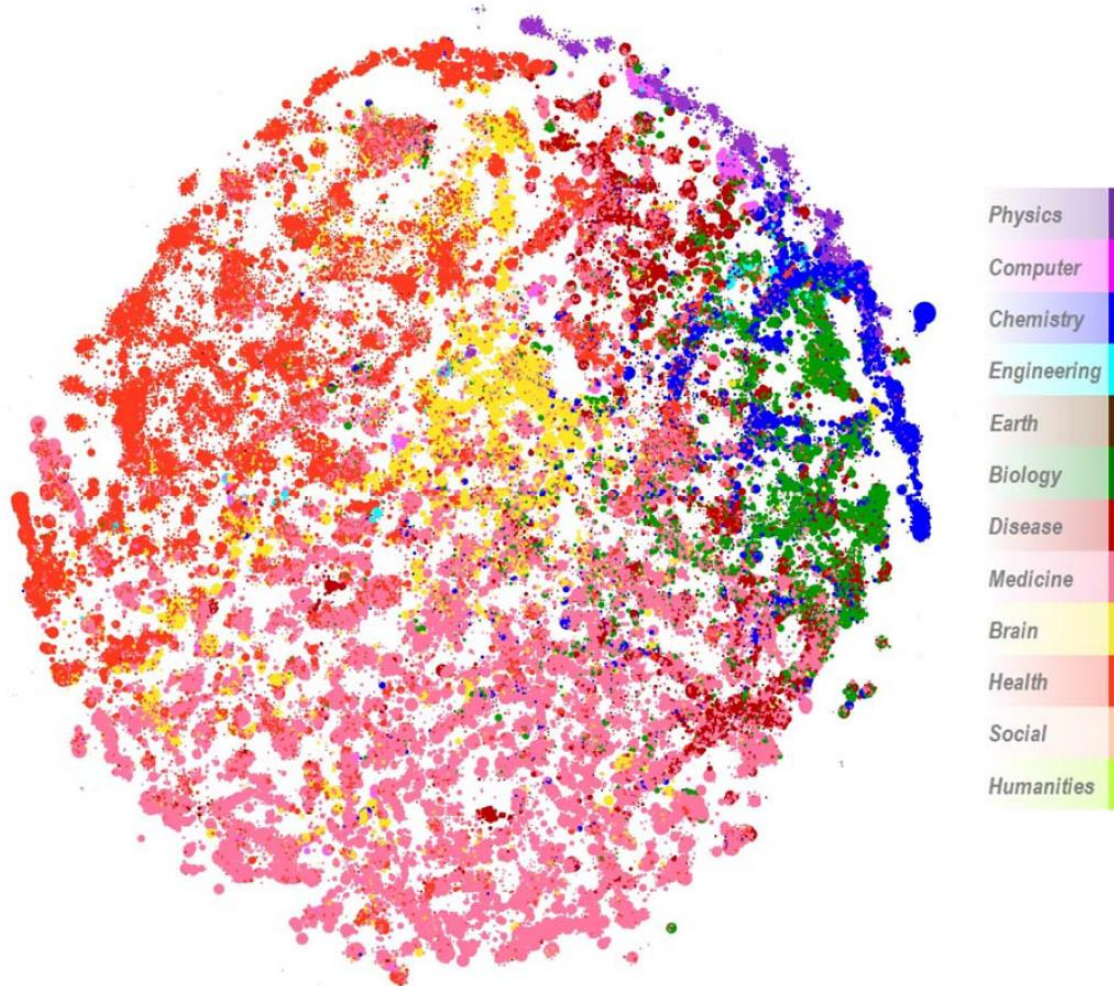
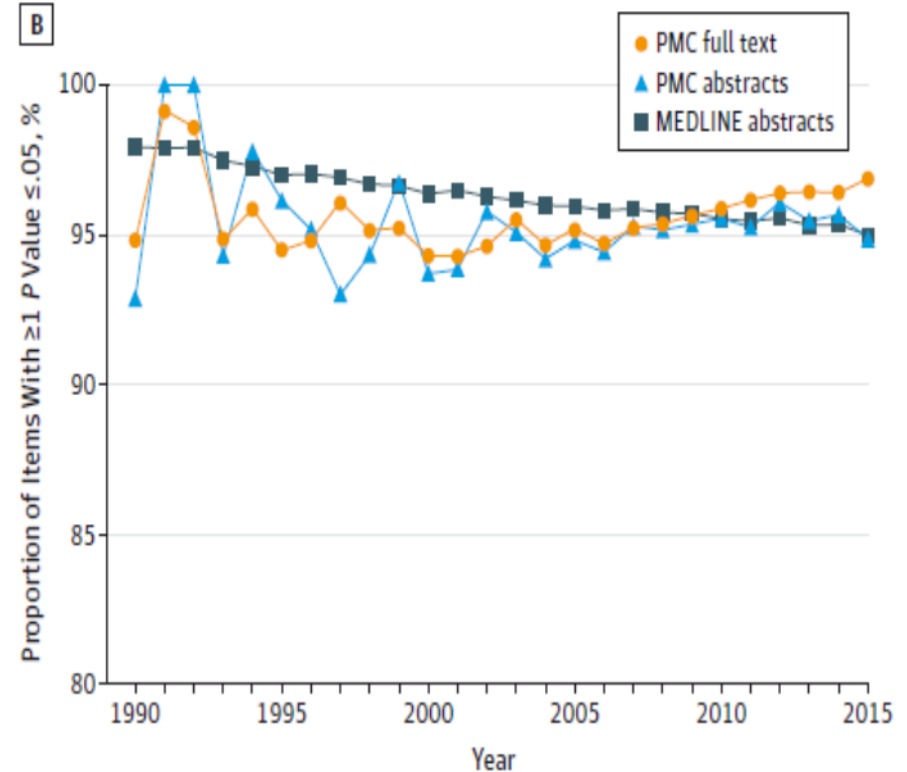
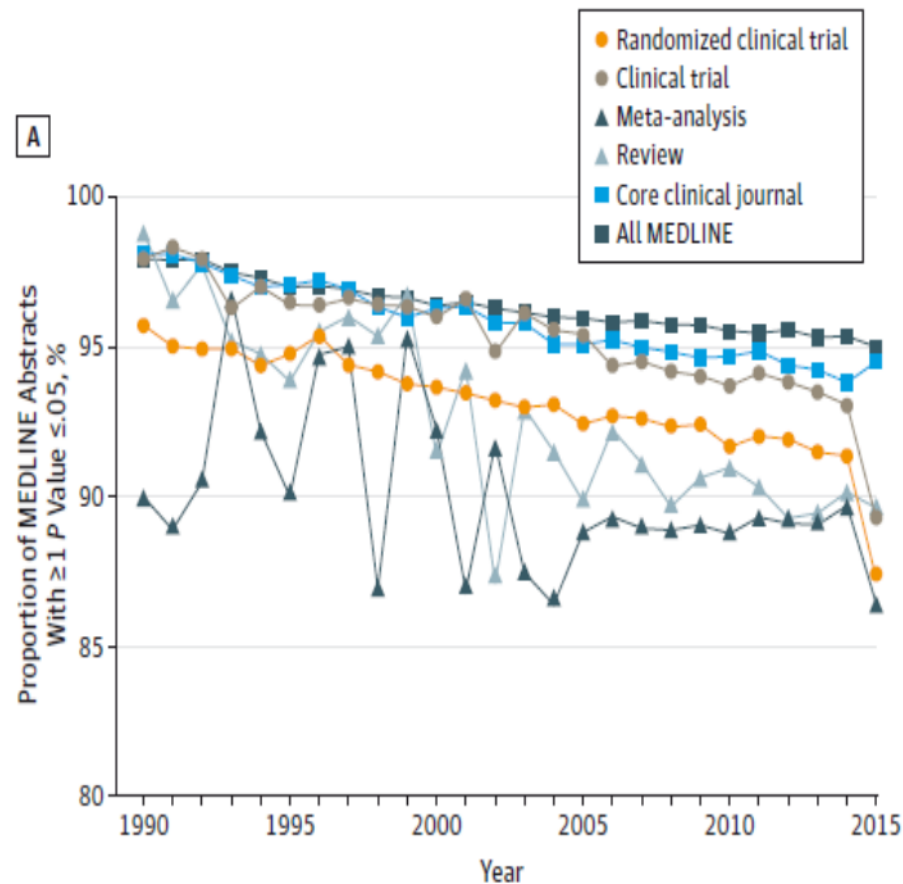














Figure 8. Visual map of the topics in the 12PM<sup>5</sup> model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic.

# Some common features apply to most fields: Statistical significance: A boring nuisance (96% of the biomedical literature claims significant results)



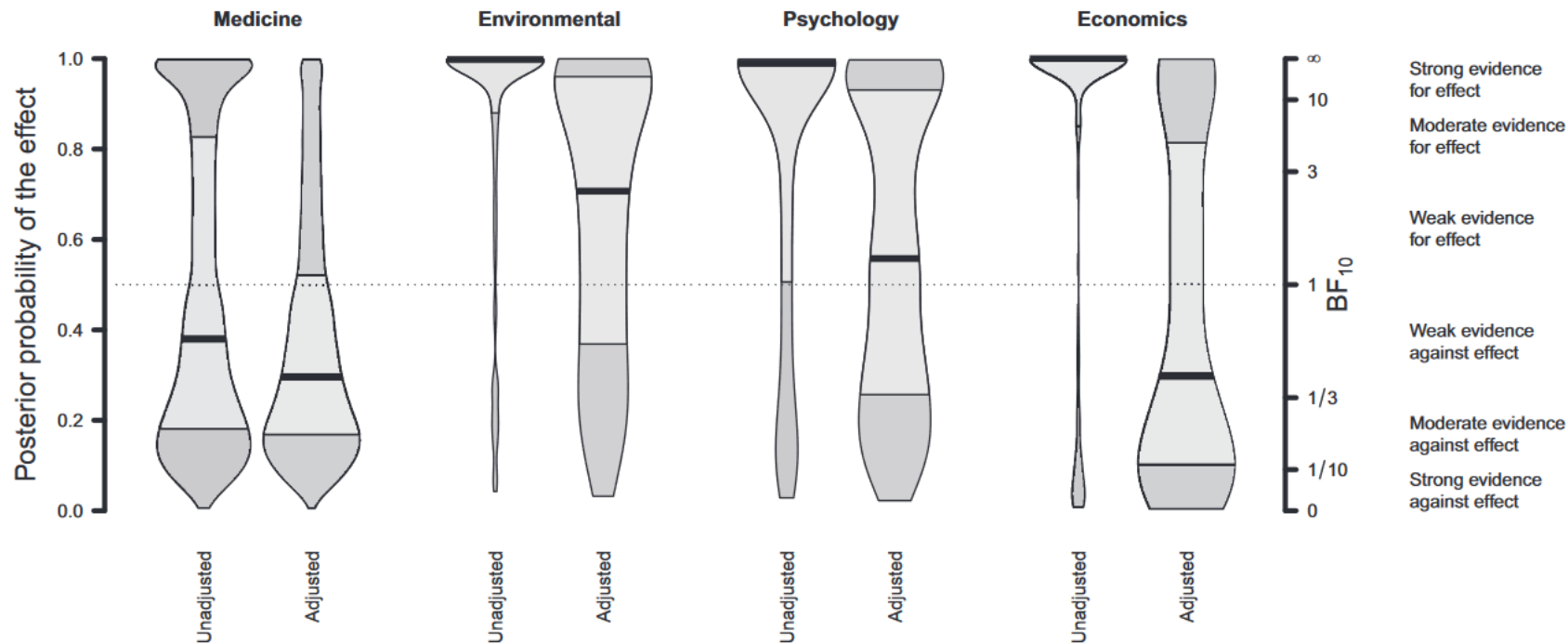
# Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics

František Bartoš<sup>1,2</sup>  | Maximilian Maier<sup>3</sup>  | Eric-Jan Wagenmakers<sup>1</sup>  |  
 Franziska Nippold<sup>4</sup>  | Hristos Doucouliagos<sup>5</sup>  | John P. A. Ioannidis<sup>6,7,8,9,10</sup>  |  
 Willem M. Otte<sup>11</sup>  | Martina Sladekova<sup>12</sup>  | Teshome K. Deressa<sup>13</sup>  |  
 Stephan B. Bruns<sup>6,13,14</sup>  | Daniele Fanelli<sup>15,16</sup>  | T. D. Stanley<sup>5</sup> 

**TABLE 1** Summary of the data sets from each field.

Field	Meta-analyses	Estimates	Estimates/MA	Effect sizes ( <i>d</i> )	Prop. significant
Medicine	67,386	597,699	5 (4, 10)	0.24 (0.09, 0.47)	0.39
Environmental	199	12,707	26 (11, 59)	0.62 (0.31, 0.95)	0.85
Psychology	605	23,563	18 (9, 40)	0.37 (0.18, 0.61)	0.78
Economics	327	91,421	66 (30, 283)	0.20 (0.09, 0.37)	0.82

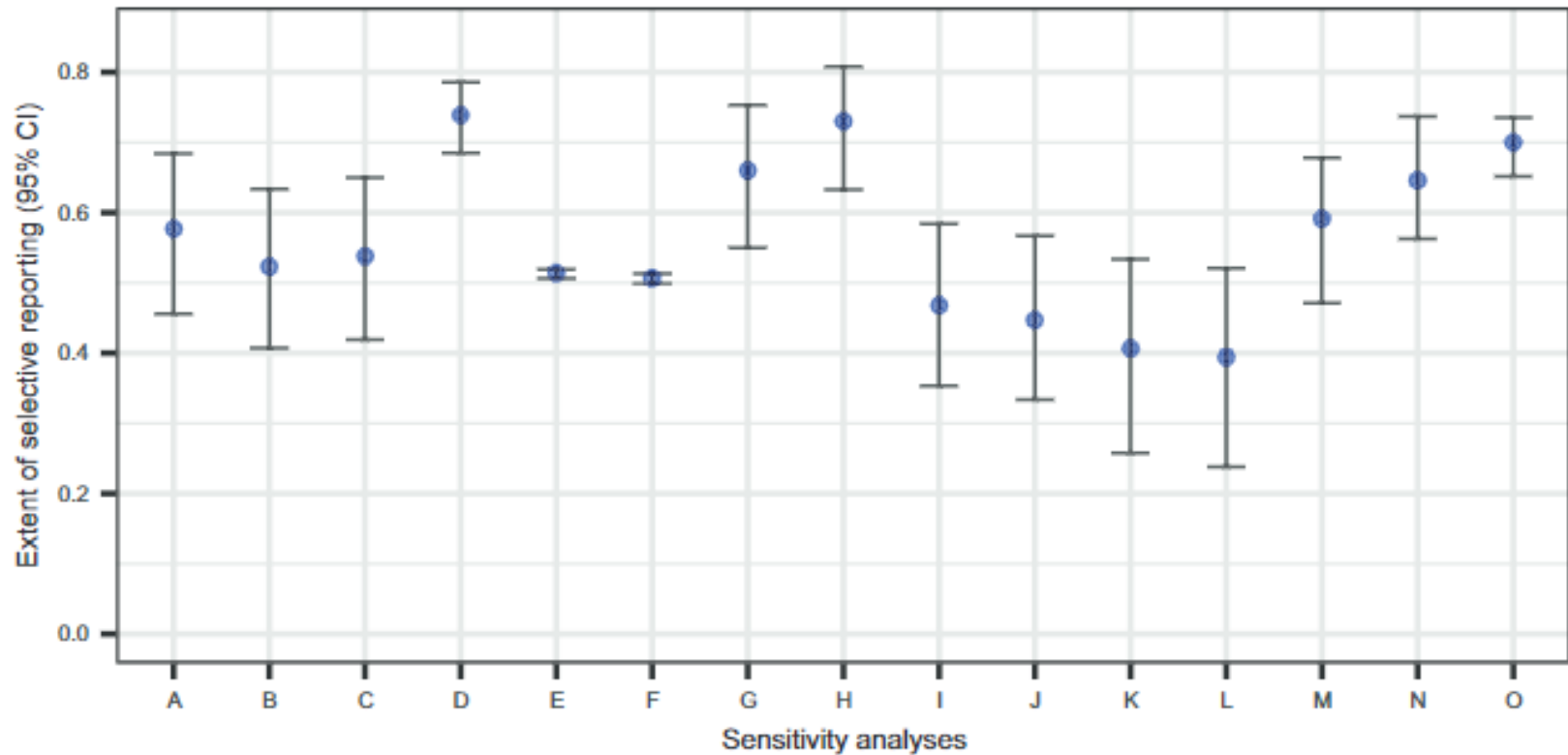




**FIGURE 1** Median, interquartile range, and distribution of posterior probability for the presence of the effect before and after adjustment for publication selection bias in each field. The width of gray area indicates density, the light gray area indicates the interquartile range, and the black line indicates the median. The y – axis is scaled according to posterior probabilities assuming equal prior probabilities of presence versus absence of the effect. See the secondary y – axis for Bayes factors in favor of the effect that are independent of the assumed prior probability of the effect.

# Estimating the extent of selective reporting: An application to economics

Stephan B. Bruns<sup>1,2,3</sup> | Teshome K. Deressa<sup>1</sup> | T. D. Stanley<sup>4</sup> |  
Chris Doucouliagos<sup>4</sup> | John P. A. Ioannidis<sup>3,5,6,7,8</sup>



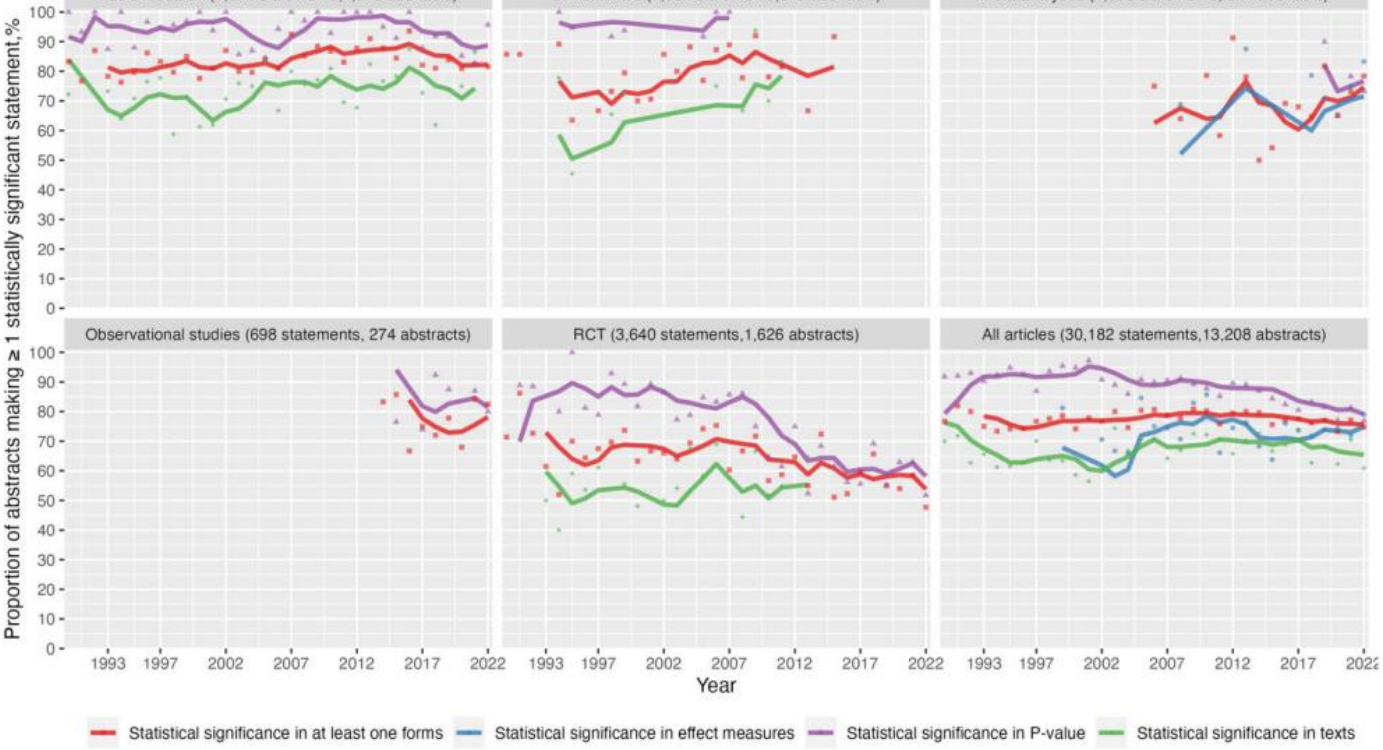
# Statistical significance becoming less common in some fields?



## Infertility

# Statistical significance and publication reporting bias in abstracts of reproductive medicine studies

Qian Feng <sup>1,\*</sup>, Ben W. Mol<sup>1,2</sup>, John P.A. Ioannidis<sup>3,4,5,6,7</sup>, and Wentao Li <sup>1</sup>



**Figure 6.** Frequency of abstracts making  $\geq 1$  statistically significant statement among abstracts making at least one statistical inference by study design. Four lines of different colours were not always present or continuous because if the total number of publications in a single year was less than six, it was not depicted. This was to show the overall trend and thus avoid the huge variations caused by a small number in a single year. The lines represent the rolling average of average proportion for 4 consecutive years, while the dots represent the exact proportion.

10.1093/humrep/dead248/7453321 by guest on 05 December 2023

Article  
TextArticle  
info

Analysis

## Inverse publication reporting bias favouring null, negative results

John P A Ioannidis <sup>1, 2, 3, 4, 5</sup>

Correspondence to Dr John P A Ioannidis, Stanford Prevention Research Center, Department of Medicine, Stanford University, 94305, USA; [jioannid@stanford.edu](mailto:jioannid@stanford.edu)

## Some examples of inverse publication reporting bias:

- Studies of toxicity and harms of interventions
- Non-inferiority studies
- Reproducibility checks?

# Modeling the scientific ecosystem: are fraud and sloppy science increasing?

PERSPECTIVE

## The credibility crisis in research: Can economics tools help?

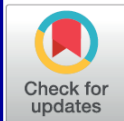
Thomas Gall<sup>1</sup>, John P. A. Ioannidis<sup>2</sup>, Zacharias Maniadis<sup>1\*</sup>

**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](mailto: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.



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

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

$$\begin{pmatrix} v_P(t) \\ v_N(t) \end{pmatrix} = \begin{pmatrix} \frac{JB}{x(t)S_{D+} + y(t)S_{C+} + z(t)S_{U+}} \\ \frac{JB}{x(t)S_{D-} + y(t)S_{C-} + z(t)S_{U-}} \end{pmatrix}.$$

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

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

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

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

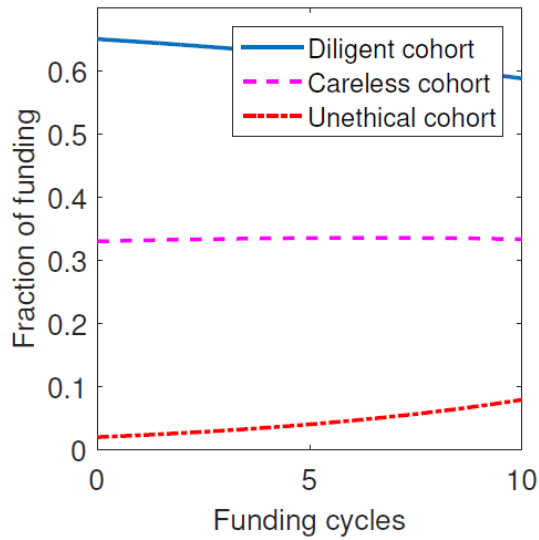
$$z(t+1) = \left( \frac{L_U(t)}{A(t)} - D_R \eta \delta v_P(t) \right) z(t) + f_U G$$

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

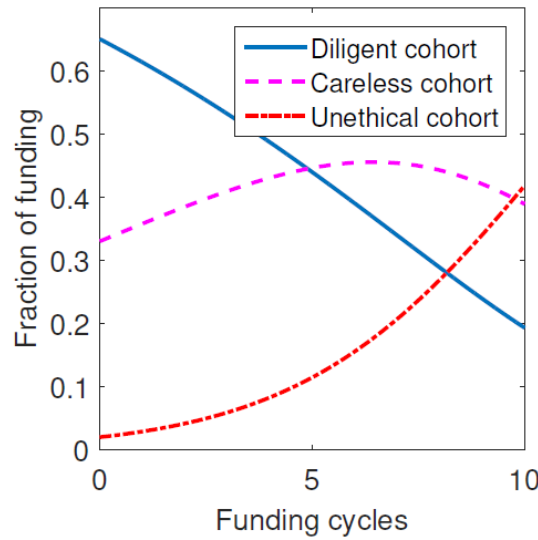
$$v(t) = \frac{J}{x(t)(S_{D+} + S_{D-}) + y(t)(S_{C+} + S_{C-}) + z(t)(S_{U+} + S_{U-})}$$

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

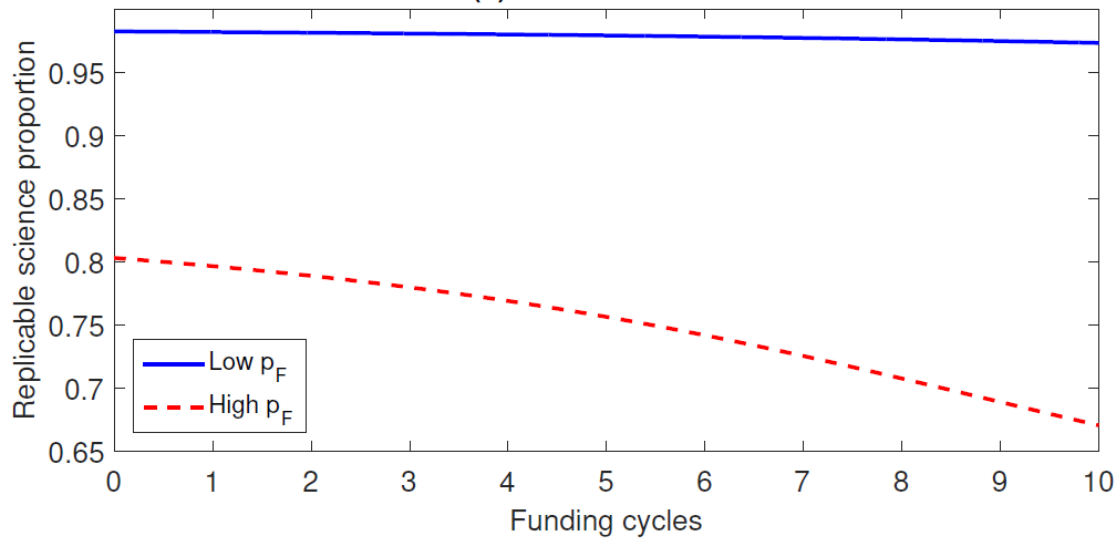
(a) Proportion of resources (Low  $p_F$ )



(b) Proportion of resources (High  $p_F$ )



(c) Trustworthiness



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

# Megajournals

- >2000 articles per year
- Acceptance rate 25-60%
- Claims for more rapid review
- Modest to large APCs

March 20, 2023

## **The Rapid Growth of Mega-Journals Threats and Opportunities**

John P. A. Ioannidis, MD, DSc<sup>1,2</sup>; Angelo Maria Pezzullo, MD, MSc<sup>3</sup>; Stefania Boccia, MSc, DSc, PhD<sup>3,4</sup>

» [Author Affiliations](#)

*JAMA*. 2023;329(15):1253-1254. doi:10.1001/jama.2023.3212





## Predatory journals: no definition, no defence

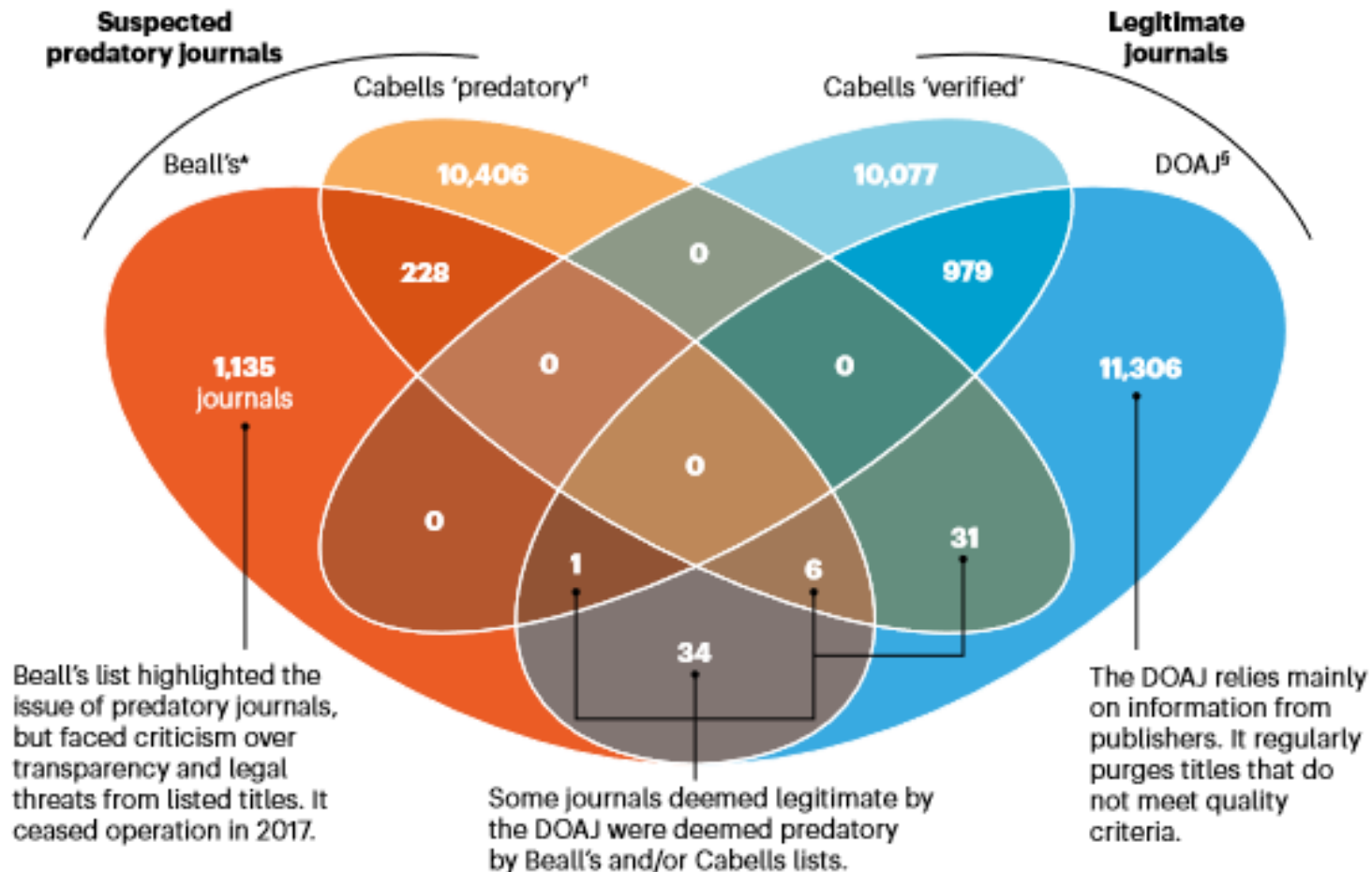
Agnes Grudniewicz, David Moher, Kelly D. Cobey and 32 co-authors

promise was doubtful and its validity unlikely to have been vetted.

Predatory journals are a global threat. They accept articles for publication – along with authors' fees – without performing promised quality checks for issues such as plagiarism or ethical approval. Naive readers are not the only victims. Many researchers have been duped into submitting to predatory journals, in which their work can be overlooked. One study that

## NO LIST TO RULE THEM ALL

Assessments of which journals are likely to be predatory or legitimate do not tally, and titles can appear in both categories. There is no way to know which journals were considered for a list but left off, or which were not considered.



<sup>a</sup>Informally assessed by University of Colorado Denver librarian Jeffrey Beall in ~2008-17; <sup>b</sup>Pay-to-access lists from Cabells, a scholarly analytics company; <sup>c</sup>The Directory of Open Access Journals, a community-curated list requiring journal best practices such as peer review and statements on author fees and licensing.

# Retracted papers originating from paper mills: cross sectional study

Cristina Candal-Pedreira,<sup>1,2</sup> Joseph S Ross,<sup>3,4,5</sup> Alberto Ruano-Ravina,<sup>1,2,6</sup> David S Egilman,<sup>7</sup> Esteve Fernández,<sup>8,9</sup> Mónica Pérez-Ríos<sup>1,2,6</sup>

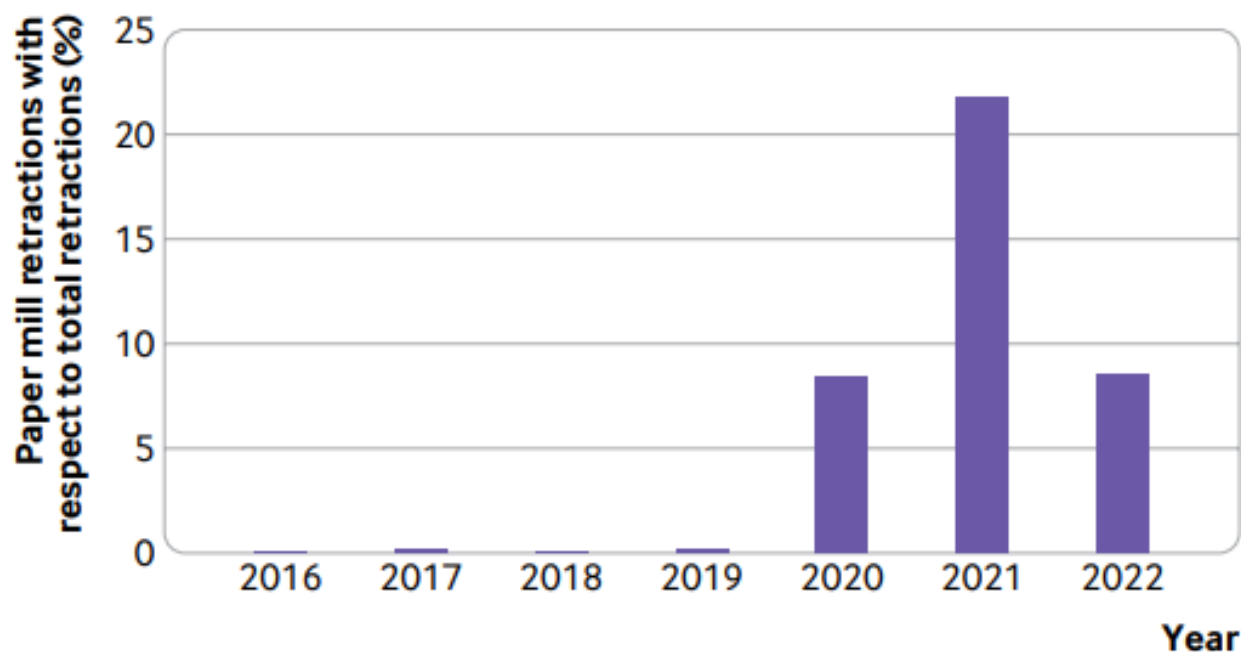


Fig 2 | Percentage of paper mill retractions with respect to total retractions

---

# AI INTENSIFIES FIGHT AGAINST PAPER MILLS

---

Text- and image-generating tools present more hurdles for efforts to tackle fake papers.

---

By **Loyal Liverpool**

researcher at New South Wales Health Pathology and the University of Sydney in Australia

# Up to one in seven submissions to hundreds of Wiley journals flagged by new paper mill tool

Wiley, whose Hindawi subsidiary has attracted thousands of paper mill papers that later needed to be retracted, has seen widespread paper mill activity among hundreds of its journals, it announced yesterday.



**Editorial**

**Hundreds of thousands of zombie randomised trials  
circulate among us**

**J. P. A. Ioannidis**

# Scientists are attracted by what is hot and gets incentivized

PNAS

RESEARCH ARTICLE

MEDICAL SCIENCES

OPEN ACCESS



## Massive covidization of research citations and the citation elite

John P. A. Ioannidis<sup>a,b,c,d,e,1</sup>, Eran Bendavid<sup>a</sup>, Maia Salholz-Hillel<sup>f</sup>, Kevin W. Boyack<sup>g</sup>, and Jeroen Baas<sup>h</sup>

Edited by Kenneth Wachter, University of California, Berkeley, CA; received March 7, 2022; accepted May 31, 2022

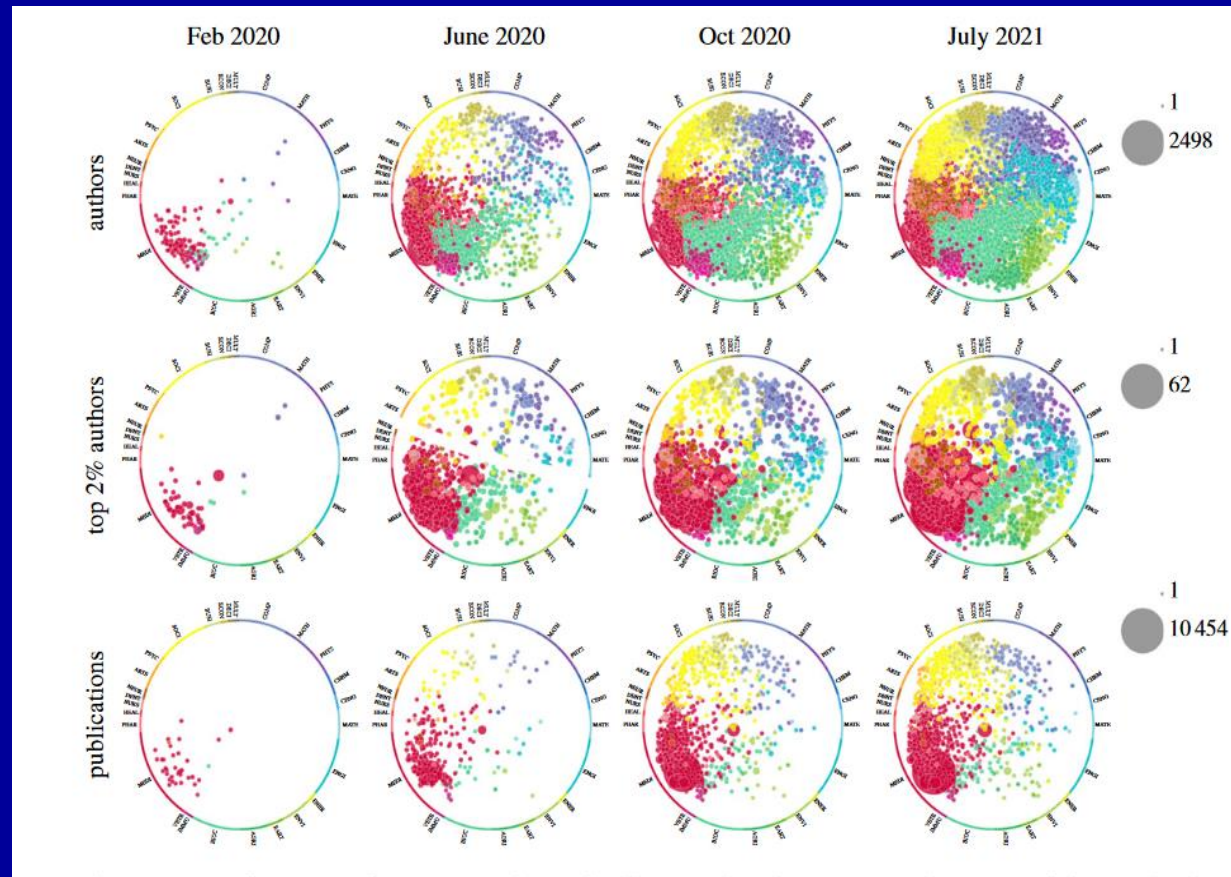
Massive scientific productivity accompanied the COVID-19 pandemic. We evaluated the citation impact of COVID-19 publications relative to all scientific work published in 2020 to 2021 and assessed the impact on scientist citation profiles. Using Scopus data until August 1, 2021, COVID-19 items accounted for 4% of papers published, 20% of citations received to papers published in 2020 to 2021, and >30% of citations received in 36 of the 174 disciplines of science (up to 79.3% in general and internal medicine). Across science, 98 of the 100 most-cited papers published in 2020 to 2021 were related to COVID-19; 110 scientists received  $\geq 10,000$  citations for COVID-19 work, but none received  $\geq 10,000$  citations for non-COVID-19 work published in 2020 to 2021. For many scientists, citations to their COVID-19 work already accounted for more than half of their total career citation count. Overall, these data show a strong covidization of research citations across science, with major impact on shaping the citation elite.

### Significance

The COVID-19 pandemic saw a massive mobilization of the scientific workforce. We evaluated the citation impact of COVID-19 publications relative to all scientific work published in 2020 to 2021, finding that 20% of citations received to papers published in 2020 to 2021 were to COVID-19-related papers. Across

~2 million scientists  
published a million  
scientific papers on  
COVID-19

(Ioannidis J. et al, Royal Society  
Open Science 2021)



**Figure 1.** Topics of prominence for COVID-19 authors and publications. The columns represent the progress of the spread at three different measuring points: by end of February 2020, end of June 2020, end of October 2020 and end of July 2021. The first row represents the spread of authors of COVID-19 papers. The authors are assigned to their most dominant topic in their career. The data are filtered to include only topics with greater than or equal to five authors assigned. The second row shows similarly the topics of the top 2% authors by field according to community detection indicators. Only topics with two or more authors are displayed. The



# Yet, quality of science suffered

## Methodological quality of COVID-19 clinical research

Richard G. Jung<sup>1,2,3,13</sup>, Pietro Di Santo<sup>1,2,4,5,13</sup>, Cole Clifford<sup>6</sup>, Graeme Proserpi-Porta<sup>7</sup>, Stephanie Skanes<sup>6</sup>, Annie Hung<sup>8</sup>, Simon Parlow<sup>4</sup>, Sarah Visintini<sup>9</sup>, F. Daniel Ramirez<sup>1,4,10,11</sup>, Trevor Simard<sup>1,2,3,4,12</sup> & Benjamin Hibbert<sup>2,3,4</sup>

The COVID-19 pandemic began in early 2020 with major health consequences. While a need to disseminate information to the medical community and general public was paramount, concerns have been raised regarding the scientific rigor in published reports. We performed a systematic review to evaluate the methodological quality of currently available COVID-19 studies compared to historical controls. A total of 9895 titles and abstracts were screened and 686 COVID-19 articles were included in the final analysis. Comparative analysis of COVID-19 to historical articles reveals a shorter time to acceptance (13.0 [IQR, 5.0-25.0] days vs. 110.0 [IQR, 71.0-156.0] days in COVID-19 and control articles, respectively;  $p < 0.0001$ ). Furthermore, methodological quality scores are lower in COVID-19 articles across all study designs. COVID-19 clinical studies have a shorter time to publication and have lower methodological quality scores than control studies in the same journal. These studies should be revisited with the emergence of stronger evidence.

## Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study

Marko Zdravkovic<sup>1</sup>, Joana Berger-Estilita<sup>2</sup>, Bogdan Zdravkovic<sup>1</sup>, David Berger<sup>3\*</sup>

## COVID-19-related medical research: a meta-research and critical appraisal



Marc Raynaud<sup>1†</sup>, Huanxi Zhang<sup>2†</sup>, Kevin Louis<sup>1†</sup>, Valentin Goutaudier<sup>1,3†</sup>, Jiali Wang<sup>2</sup>, Quentin Dubourg<sup>4</sup>, Yongcheng Wei<sup>2</sup>, Zeynep Demir<sup>1,5</sup>, Charlotte Debais<sup>1</sup>, Olivier Aubert<sup>1</sup>, Yassine Bouatou<sup>1</sup>, Carmen Lefaucheur<sup>6</sup>, Patricia Jabre<sup>7</sup>, Longshan Liu<sup>2</sup>, Changxi Wang<sup>2</sup>, Xavier Jouven<sup>1</sup>, Peter Reese<sup>1,8</sup>, Jean-Philippe Empana<sup>1</sup> and Alexandre Loupy<sup>1\*</sup>

### Abstract

**Background:** Since the start of the COVID-19 outbreak, a large number of COVID-19-related papers have been published. However, concerns about the risk of expedited science have been raised. We aimed at reviewing and categorizing COVID-19-related medical research and to critically appraise peer-reviewed original articles.

**Methods:** The data sources were Pubmed, Cochrane COVID-19 register study, arXiv, medRxiv and bioRxiv, from 01/11/2019 to 01/05/2020. Peer-reviewed and preprints publications related to COVID-19 were included, written in English or Chinese. No limitations were placed on study design. Reviewers screened and categorized studies according to *i*) publication type, *ii*) country of publication, and *iii*) topics covered. Original articles were critically appraised using validated quality assessment tools.

**Results:** Among the 11,452 publications identified, 10,516 met the inclusion criteria, among which 7468 (71.0%) were peer-reviewed articles. Among these, 4190 publications (56.1%) did not include any data or analytics (comprising expert opinion pieces). Overall, the most represented topics were infectious disease ( $n = 2326$ , 22.1%), epidemiology ( $n = 1802$ , 17.1%), and global health ( $n = 1602$ , 15.2%). The top five publishing countries were China (25.8%), United States (22.3%), United Kingdom (8.8%), Italy (8.1%) and India (3.4%). The dynamic of publication showed that the exponential growth of COVID-19 peer-reviewed articles was mainly driven by publications without original data (mean 261.5 articles  $\pm$  51.1 per week) as compared with original articles (mean of 69.3  $\pm$  22.3 articles per week). Original articles including patient data accounted for 713 (9.5%) of peer-reviewed studies. A total of 576 original articles (80.8%) showed intermediate to high risk of bias. Last, except for simulation studies that mainly used large-scale open data, the median number of patients enrolled was of 102 (IQR = 37–337).

**Conclusions:** Since the beginning of the COVID-19 pandemic, the majority of research is composed by publications without original data. Peer-reviewed original articles with data showed a high risk of bias and included a limited number of patients. Together, these findings underscore the urgent need to strike a balance between the velocity and quality of research, and to cautiously consider medical information and clinical applicability in a pressing, pandemic context.

(Continued on next page)

# Do we need revolution or simply evolution?

## A manifesto for reproducible science

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

Improving the reliability and efficiency of scientific research will increase the credibility of the published scientific literature and accelerate discovery. Here we argue for the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives. There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures, but their broad adoption by researchers, institutions, funders and journals will require iterative evaluation and improvement. We discuss the goals of these measures, and how they can be implemented, in the hope that this will facilitate action toward improving the transparency, reproducibility and efficiency of scientific research.

Identify problems or push for solutions?

# Why Most Published Research Findings Are False

OPEN  ACCESS Freely available online

 **PLOS** | MEDICINE

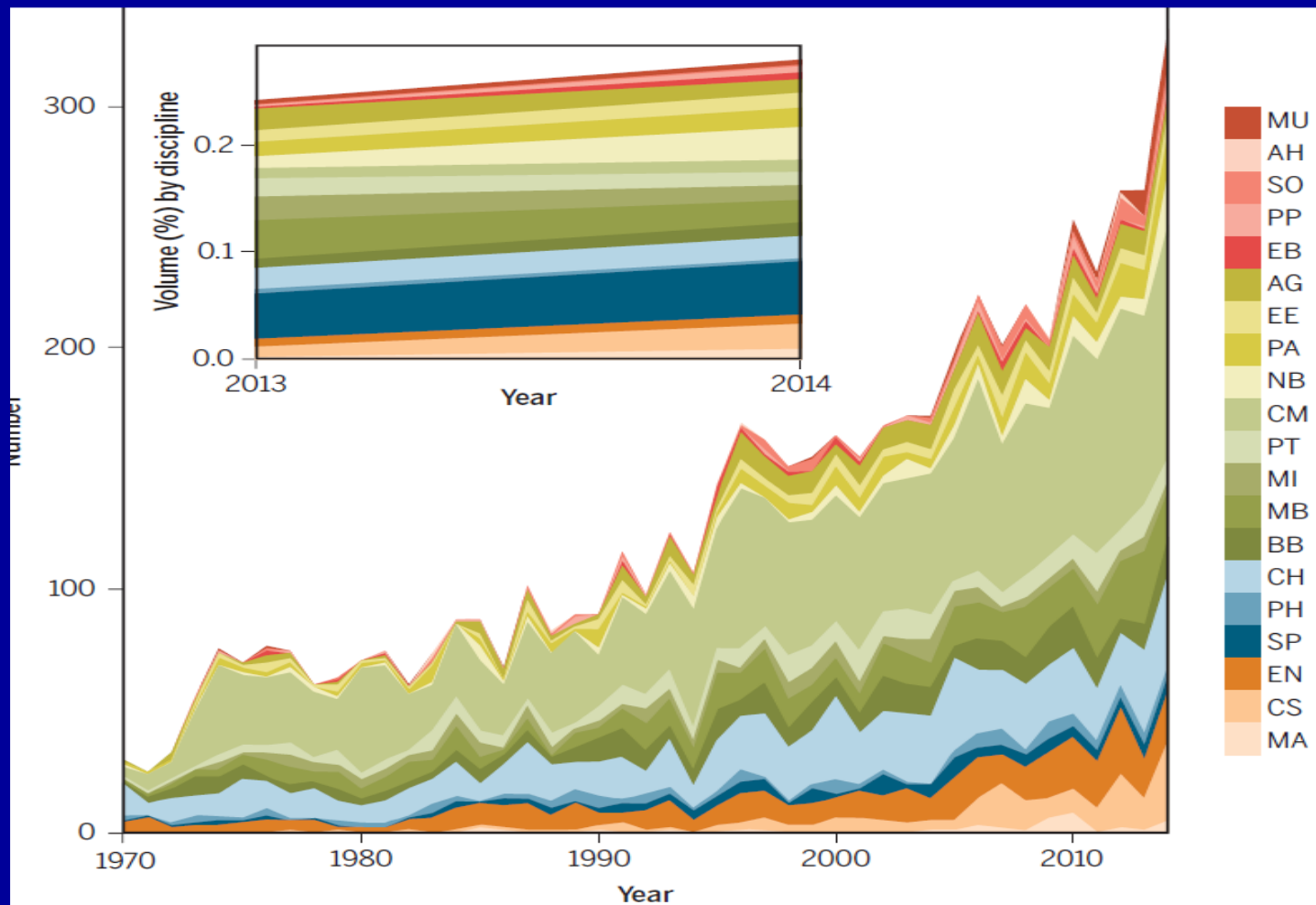
Essay

How to Make More Published Research True

# 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.”



# 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.

# Improvements in reproducibility

- Reproducibility of methods: yes, in some fields, but not necessarily in those that produce many papers
- Reproducibility of results: remains unknown in many/most fields and most papers where replication is not attempted
- Reproducibility of inferences: is it even possible (worthwhile?) to improve

# Inferential reproducibility may be doomed to be modest (or low) by its very nature

## JAMA Health Forum™

---

### Viewpoint

## The Subjective Interpretation of the Medical Evidence

Howard Bauchner, MD; John P. A. Ioannidis, MD, DSc

---

Experts often subjectively disagree on how they interpret the same evidence and what recommendations they derive from it.<sup>1</sup> Meticulous processes to resolve diverging views in guideline development efforts, for example, may not remove subjectivity. Even the most prestigious organizations sometimes have different guideline recommendations. Subjective disagreements can be common, extreme, and unsettling when evidence is limited and rapidly evolving—as in many questions related to COVID-19. However, subjectivity exists, and differences ensue even for common diseases where evidence has accrued and been evaluated for decades. For example, the American College of Physicians, the American Cancer Society, and the US Preventive Services Task Force (USPSTF) vary on when to initiate screening for colorectal cancer and the preferred screening methods.<sup>2-4</sup> Breast cancer and depression screening recommendations have been debated for decades.

# 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

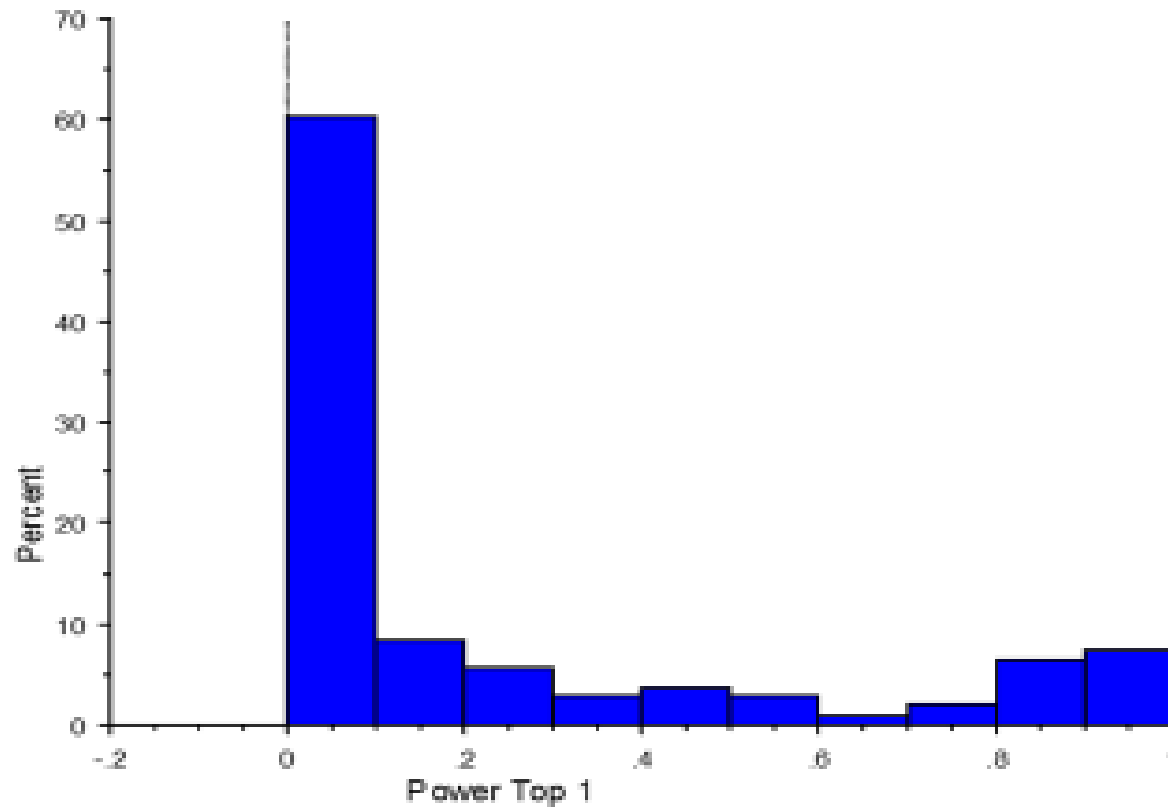


# Power failure: why small sample size undermines the reliability of neuroscience

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

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

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



# 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

# Big Data, Big Noise, Big Error

## MEDICINE

### *Big data meets public health*

Human well-being could benefit from large-scale data if large-scale noise is minimized

By **Muin J. Khoury<sup>1</sup>** and  
**John P. A. Ioannidis<sup>2</sup>**

**N**1854, as cholera swept through London, John Snow, the father of modern epidemiology, painstakingly recorded the locations of affected homes. After long, laborious work, he implicated the Broad Street water pump as the source of the outbreak, even without knowing that a *Vibrio* organism caused cholera. Today, Snow might have crunched Global Positioning System information and disease prevalence data, solving the problem within hours (1). That is the potential impact of “Big Data” on the public’s health. But the promise of Big Data is also accompanied by claims that “the scientific method itself is becoming obsolete” (2), as next generation computers, such as IBM’s Watson (3), sift through the digital world to provide predictive models based on massive information. Separating the true signal from the gigantic amount of noise is neither easy nor straightforward, but it is a challenge that must be tackled if information is ever to be translated into societal well being.

The term “Big Data” refers to volumes of large, complex, linkable information (4). Beyond genomics and other “omic” fields, Big Data includes medical, environmental, financial, geographic, and social media information. Most of this digital information was unavailable a decade ago. This swell of data will continue to grow, stoked by sources that are currently unimaginable. Big Data stands to improve health by providing insights into



**From validity to utility.** Big Data can improve tracking and response to infectious disease outbreaks, discovery of early warning signals of disease, and development of diagnostic tests and therapeutics.

For non-genomic associations, false alarms due to confounding variables or other biases are possible even with very large-scale studies, extensive replication, and very strong signals (9). Big Data’s strength is in finding associations, not in showing whether these associations have meaning. Finding a signal is only the first step.

Even John Snow needed to start with a plausible hypothesis to know where to look, i.e., choose what data to examine. If all he had was massive amounts of data, he might well have ended up with a correlation as spurious as the honey bee-marijuana connection. Crucially, Snow “did the experiment.” He removed the handle from the water pump and dramatically reduced the spread of cholera, thus moving from correlation to causation and effective intervention.

How can we improve the potential for Big Data to improve health and prevent disease? One priority is that a stronger epidemiological foundation is needed. Big Data analysis is currently largely based on convenient samples of people or information available on the Internet. When associations are probed between perfectly measured data (e.g., a genome sequence) and poorly measured data (e.g., administrative claims health data), research accuracy is dictated by the weakest link. Big Data are observational in nature and are fraught with many biases such as selection, confounding variables, and lack of generalizability. Big Data analysis may be embedded in epidemiologically well-characterized and representative populations. This epi-

# Small data, big data, no data

VIEWPOINT

## Stealth Research

Is Biomedical Innovation Happening Outside the Peer-Reviewed Literature?

Ioannidis, JAMA, 2015

## Stealth research: Lack of peer-reviewed evidence from healthcare unicorns

Ioana A. Cristea<sup>1,2</sup> | Eli M. Cahan<sup>3,4</sup> | John P. A. Ioannidis<sup>1,5,6,7,8</sup>

### Key messages

- Start-ups are widely accepted as key vehicles of innovation and disruption in healthcare, positioned to make revolutionary discoveries.
- Most of the highest-valued start-ups in healthcare have a limited or non-existent participation and impact in the publicly available scientific literature.
- The system of peer-reviewed publishing, while imperfect, is indispensable for validating innovative products and technologies in biomedicine.
- Healthcare products not subjected to peer-review but based on internal data generation alone may be problematic and non-trustworthy.

# AI and the increasing dark matter of research production

- Stanford has the highest computational capacity than any other university worldwide
- Still, this is only 1% of the computational capacity of Alphabet, Meta, Microsoft, Apple.
- Most AI research may grow outside the (published) scientific literature
- It may or may not be open source

# Investigating the replicability of preclinical cancer biology

Timothy M Errington<sup>1\*</sup>, Maya Mathur<sup>2</sup>, Courtney K Soderberg<sup>1</sup>,  
Alexandria Denis<sup>1†</sup>, Nicole Perfito<sup>1‡</sup>, Elizabeth Iorns<sup>3</sup>, Brian A Nosek<sup>1,4</sup>

<sup>1</sup>Center for Open Science, Charlottesville, United States; <sup>2</sup>Quantitative Sciences Unit, Stanford University, Stanford, United States; <sup>3</sup>Science Exchange, Palo Alto, United States; <sup>4</sup>University of Virginia, Charlottesville, United States

---

**Abstract** Replicability is an important feature of scientific research, but aspects of contemporary research culture, such as an emphasis on novelty, can make replicability seem less important than it should be. The [Reproducibility Project: Cancer Biology](#) was set up to provide evidence about the replicability of preclinical research in cancer biology by repeating selected experiments from high-impact papers. A total of 50 experiments from 23 papers were repeated, generating data about the replicability of a total of 158 effects. Most of the original effects were positive effects (136), with the rest being null effects (22). A majority of the original effect sizes were reported as numerical values (117), with the rest being reported as representative images (41). We employed seven methods to assess replicability, and some of these methods were not suitable for all the effects in our sample. One method compared effect sizes: for positive effects, the median effect size in the replications was 85% smaller than the median effect size in the original experiments, and 92% of replication effect sizes were smaller than the original. The other methods were binary – the replication was either a success or a failure – and five of these methods could be used to assess both positive and null effects when effect sizes were reported as numerical values. For positive effects, 40% of replications (39/97) succeeded according to three or more of these five methods, and for null effects 80% of replications (12/15) were successful on this basis; combining positive and null effects, the success rate was 46% (51/112). A successful replication does not definitively confirm an original finding or its theoretical interpretation. Equally, a failure to replicate does not disconfirm a finding, but it does suggest that additional investigation is needed to establish its reliability.

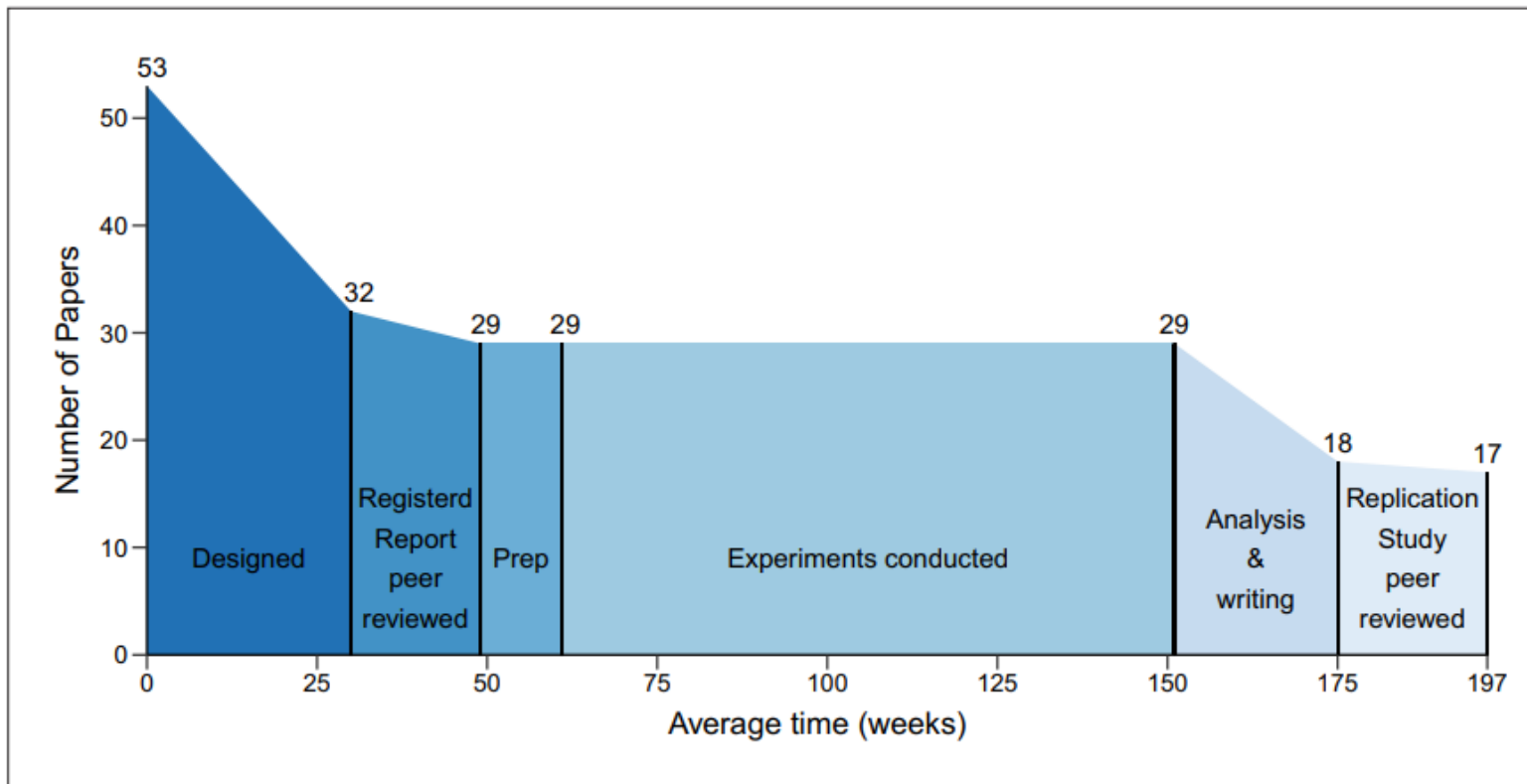


## REPRODUCIBILITY IN CANCER BIOLOGY

# Challenges for assessing replicability in preclinical cancer biology

**Abstract** We conducted the [Reproducibility Project: Cancer Biology](#) to investigate the replicability of preclinical research in cancer biology. The initial aim of the project was to repeat 193 experiments from 53 high-impact papers, using an approach in which the experimental protocols and plans for data analysis had to be peer reviewed and accepted for publication before experimental work could begin. However, the various barriers and challenges we encountered while designing and conducting the experiments meant that we were only able to repeat 50 experiments from 23 papers. Here we report these barriers and challenges. First, many original papers failed to report key descriptive and inferential statistics: the data needed to compute effect sizes and conduct power analyses was publicly accessible for just 4 of 193 experiments. Moreover, despite contacting the authors of the original papers, we were unable to obtain these data for 68% of the experiments. Second, none of the 193 experiments were described in sufficient detail in the original paper to enable us to design protocols to repeat the experiments, so we had to seek clarifications from the original authors. While authors were *extremely* or *very helpful* for 41% of experiments, they were *minimally helpful* for 9% of experiments, and *not at all helpful* (or did not respond to us) for 32% of experiments. Third, once experimental work started, 67% of the peer-reviewed protocols required modifications to complete the research and just 41% of those modifications could be implemented. Cumulatively, these three factors limited the number of experiments that could be repeated. This experience draws attention to a basic and fundamental concern about replication – it is hard to assess whether reported findings are credible.

**TIMOTHY M ERRINGTON\***, **ALEXANDRIA DENIS<sup>†</sup>**, **NICOLE PERFITO<sup>‡</sup>**,  
**ELIZABETH IORNS AND BRIAN A NOSEK**



**Figure 4.** The different phases of the replication process. Graph showing the number of papers entering each of the six phases of the replication process, and the mean duration of each phase in weeks. 53 papers entered the design phase, which started with the selection of papers for replication and ended with submission of a Registered Report (mean = 30 weeks; median = 31; IQR = 21–37). 32 papers entered the protocol peer reviewed phase, which ended with the acceptance of a Registered Report (mean = 19 weeks; median = 18; IQR = 15–24). 29 papers entered the preparation phase (Prep), which ended when experimental work began (mean = 12 weeks; median = 3; IQR = 0–11). The mean for the prep phase was much higher than the median (and outside the IQR) because this phase took less than a week for many studies, but much longer for a small number of studies. The same 29 papers entered the conducted phase, which ended when the final experimental data were delivered (mean = 90 weeks; median = 88; IQR = 44–127), and the analysis and writing phase started, which ended with the submission of a Replication Study (mean = 24 weeks; median = 23; IQR = 7–32). 18 papers entered the results peer review phase, which ended with the acceptance of a Replication Study (mean = 22 weeks; median = 18; IQR = 15–26). In the end, 17 Replication Studies were accepted for publication. The entire process had a mean length of 197 weeks and a median length of 181 weeks (IQR = 102–257).

Clinical Chemistry 63:5  
000-000 (2017)

Perspectives

---

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

John P.A. Ioannidis<sup>1,2,3,4\*</sup>

# Resistance to refutation

## Persistence of Contradicted Claims in the Literature

Athina Tatsioni, MD  
Nikolaos G. Bonitsis, MD  
John P. A. Ioannidis, MD

SOME RESEARCH FINDINGS THAT have received wide attention in the scientific community, as proven by the high citation counts of the respective articles, are eventually contradicted by subsequent evidence.<sup>1</sup> A number of such high-profile contradictions pertain to differences between nonrandomized and randomized studies. For example, the effect of vitamin E on cardiovascular disease prevention has been in the center of a major debate in clinical research over the last 2 decades. Vitamin E is known to have antioxidant activity, and a long list of citations in the preclinical literature on antioxidants<sup>2-4</sup> suggested that these agents may be beneficial for cancer and cardiovascular disease. Two highly cited publications suggested in the 1990s that vitamin E could decrease cardiovascular disease risk by almost half in men and in women.<sup>5,6</sup> However, subsequent randomized trials showed no benefit or even suggested increased harm.<sup>7,8</sup> Several other highly prominent contradictions have also been recorded pertaining to the effects of other dietary components and hormones.<sup>9-15</sup> The prominent refutation of the epidemiological studies has spurred considerable controversy for observational epidemiology in general.<sup>16-21</sup>

Such debate offers opportunities to study what happens to the scientific literature, when a highly prominent claim is refuted. How quickly are such beliefs abandoned? Is there still literature citing the contradicted studies despite their refutation? What counterarguments are

**Context** Some research findings based on observational evidence dictated by randomized trials, but may nevertheless still be supported by circles.

**Objectives** To evaluate the change over time in the content of cited epidemiological studies that proposed major cardiovascular benefits of vitamin E in 1993; and to understand how these benefits in the literature, despite strong contradicting evidence from large randomized trials (RCTs). To examine the generalizability of these findings, we of persistence of supporting citations for the highly cited and effects of beta-carotene on cancer and of estrogen on Alzheimer

**Data Sources** For vitamin E, we sampled articles published (before, early, and late after publication of refuting evidence) cited epidemiological studies and separately sampled article referencing the major contradicting RCT (HOPE trial). We a lished in 2006 that referenced highly cited articles proposing beta-carotene for cancer (published in 1981 and contradic 1994-1996) and estrogen for Alzheimer disease (published i recently by RCTs in 2004).

**Data Extraction** The stance of the citing articles was rated and unfavorable to the intervention. We also recorded th ements raised to defend effectiveness against contradicting e

**Results** For the 2 vitamin E epidemiological studies, even i ticles remained favorable. A favorable stance was independe cent articles, specifically in articles that also cited the HOPE trial [95% confidence interval, 0.01-0.19;  $P < .001$ ] and the odds i confidence interval, 0.02-0.24;  $P < .001$ ], as compared with internal medicine vs specialty journals. Among articles citing the l were unfavorable. In 2006, 62.5% of articles referencing the h proposed beta-carotene and 61.7% of those referencing the highly cited article on estrogen effectiveness were still favorable; 100% and 96%, respectively, of the citations appeared in specialty journals; and citations were significantly less favorable ( $P = .001$  and  $P = .009$ , respectively) when the major contradicting trials were also mentioned. Counterarguments defending vitamin E or estrogen included diverse selection and information biases and genuine differences across studies in participants, interventions, counter-ventions, and outcomes. Favorable citations to beta-carotene, long after evidence contradicted its effectiveness, did not consider the contradicting evidence.

**Conclusion** Claims from highly cited observational studies persist and continue to be supported in the medical literature despite strong contradictory evidence from randomized trials.

JAMA. 2007;298(21):2517-2526

www.jama.com

**Author Affiliations:** Department of Hygiene and Epidemiology, (Drs Tatsioni, Bonitsis, and Ioannidis) and the Department of Dermatology (Dr Bonitsis), University of Ioannina School of Medicine; and the Biomedical Research Institute, Foundation for Research and Technology-Hellas (Dr Ioannidis), Ioannina, Greece; Institute for Clinical Research and Health Policy

Studies, Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts (Drs Tatsioni and Ioannidis).

**Corresponding Author:** John P. A. Ioannidis, MD, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, University Campus, Ioannina, 45110 Greece (jioannid@cc.uoi.gr).

Advances in Methods and Practices in Psychological Science

Volume 4, Issue 3, July-September 2021




© The Author(s) 2021, Article Reuse Guidelines

<https://doi.org/10.1177/25152459211040837>



### Empirical Article

## Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology

Tom E. Hardwicke <sup>1,2</sup>, Dénes Szűcs<sup>3</sup>, Robert T. Thibault <sup>4,5</sup>, Sophia Crüwell <sup>2,6</sup>, Olmo R. van den Akker<sup>7</sup>, Michèle B. Nuijten<sup>7</sup>, and John P. A. Ioannidis<sup>2,8,9</sup>

## **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

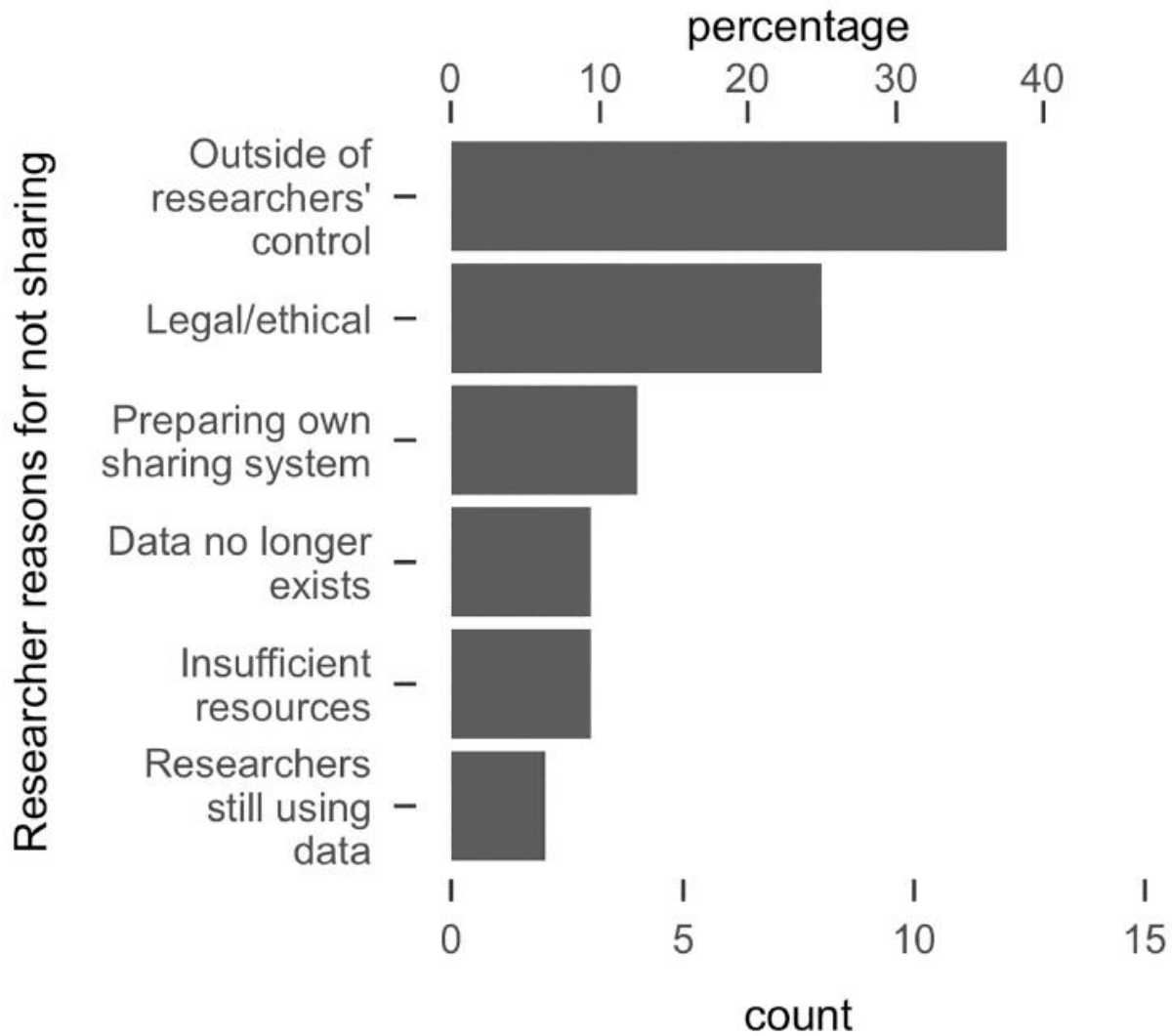
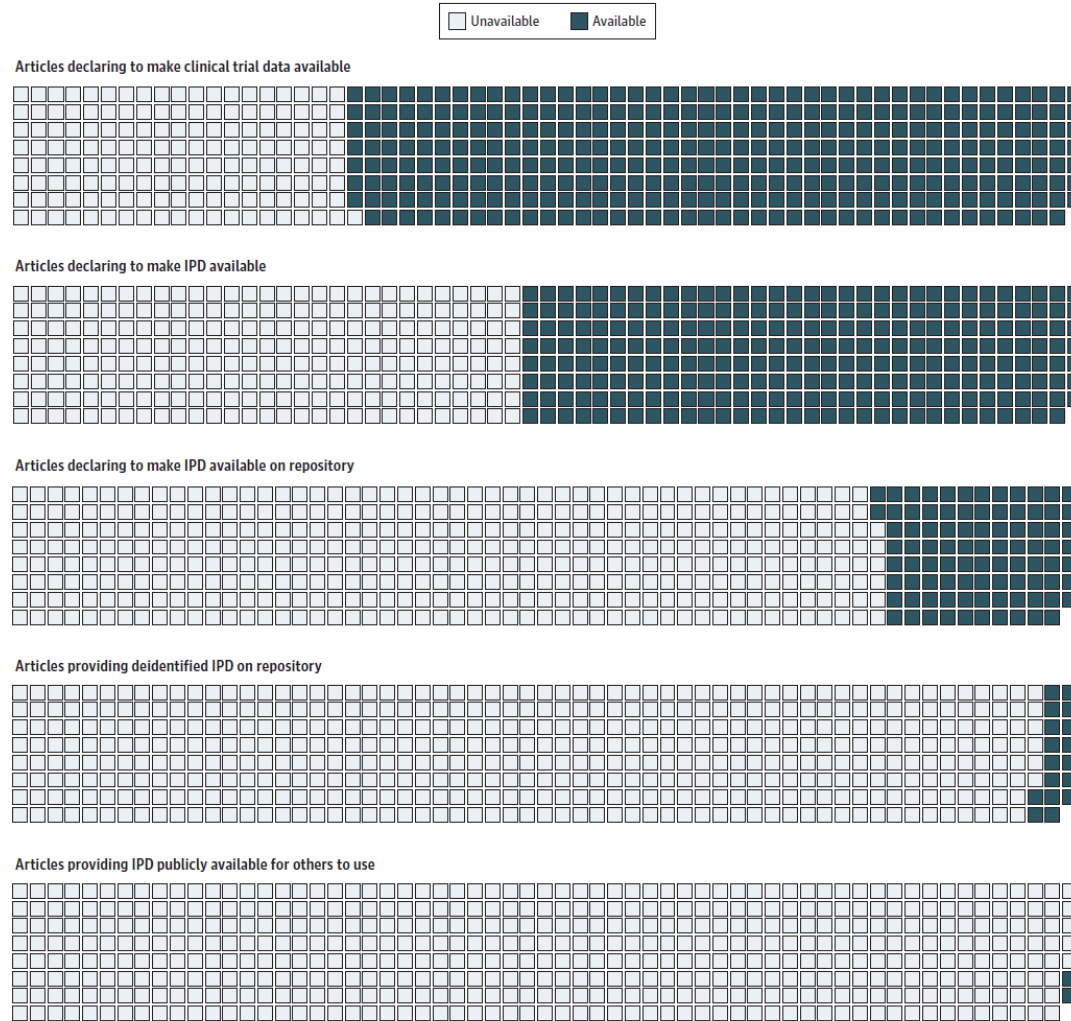


Fig 2. Reasons provided by researchers for not sharing. X-axes represent counts and percentages (of n = 32 who responded that they would not share).

# Evaluation of Data Sharing After Implementation of the International Committee of Medical Journal Editors Data Sharing Statement Requirement

Valentin Danchev, DPhil; Yan Min, MD; John Borghi, PhD; Mike Baiocchi, PhD; John P. A. Ioannidis, MD, DSc

Figure 3. Indicators of Declared and Actual Clinical Trial Individual-Participant Data (IPD) Availability as of April 10, 2020



# Lifting of Embargoes to Data Sharing in Clinical Trials Published in Top Medical Journals

Maximilian Siebert, PhD<sup>1</sup>; John P. A. Ioannidis, MD, DSc<sup>1</sup>

**Table 1. Prevalence and Conditions of Data Sharing Practices and Mechanisms Across the Included Trials**

Table 1. Prevalence and Conditions of Data Sharing Practices and Mechanisms Across the Included Trials

Studies overview	No./total No. (%)				
	Total (N = 158)	Industry sponsoring (n = 49)	NIH sponsoring (n = 22)	Nonindustry and non-NIH (n = 57)	Mixed sponsoring (n = 30)
<b>Embargo lifting<sup>a</sup></b>					
Lifted data sharing embargo	104/158 (65.8)	31/49 (63.3)	18/22 (81.8)	40/57 (70.2)	15/30 (50)
No reply after 3 reminders	42/158 (26.6)	11/49 (22.4)	3/22 (13.6)	15/57 (26.3)	13/30 (43.3)
Refused data sharing	12/158 (7.6)	7/49 (14.3)	1/22 (4.6)	2/57 (3.5)	2/30 (6.7)
<b>Data sharing mechanism</b>					
Data repositories (eg, Vivli, NIH databases) <sup>b</sup>	48/104 (46.2)	21/31 (67.8)	12/18 (66.7)	6/40 (15)	9/15 (60)
Direct request to authors	29/104 (27.9)	No trials	5/18 (27.8)	20/40 (50)	4/15 (26.6)
Company requests	9/104 (8.6)	9/31 (29)	No trials	No trials	No trials
Requests to groups, committees, or units	6/104 (5.8)	No trials	1/18 (5.5)	4/40 (10)	1/15 (6.7)
Others (eg, OneDrive plus email)	4/104 (3.8)	No trials	No trials	4/40 (10)	No trials
Mechanism unclear	8/104 (7.7)	1/31 (3.2)	No trials	6/40 (15)	1/15 (6.7)
<b>Mechanism change</b>					
Consistent with original statement	77/104 (74)	23/31 (74.2)	15/18 (83.3)	28/40 (70)	11/15 (73.3)
Mechanism unclear	8/104 (7.7)	1/31 (3.2)	No trials	6/40 (15)	1/15 (6.7)
Different from original statement	19/104 (18.3)	7/31 (22.6)	3/18 (16.7)	6/40 (15)	3/15 (20)

Abbreviation: NIH, National Institutes of Health.

<sup>a</sup> Two data sets still had active embargoes when our survey commenced. One was set to expire in September 2023, and we received no response regarding it, while the other was set for October 2024, and we located the data set on Vivli.

<sup>b</sup> Four of those trials had data that were freely available without requiring research proposal submission.



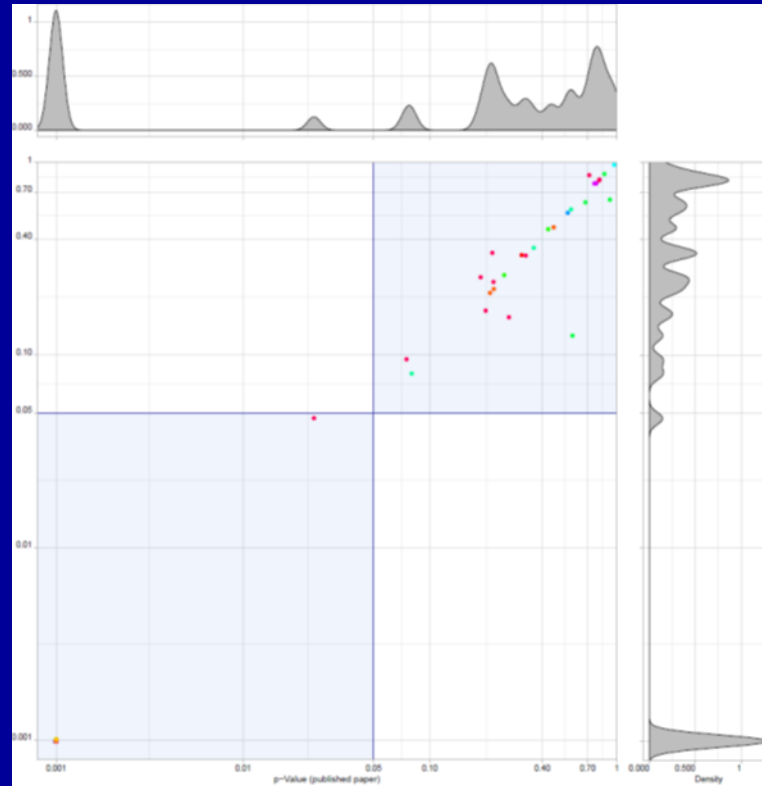
# Industry Involvement and Transparency in the Most Cited Clinical Trials, 2019-2022

Leonardo M. Siena, MD; Lazaros Papamanolis, MSc; Maximilian J. Siebert, PhD; Rosa Katia Bellomo, MD; John P. A. Ioannidis, MD, DSc

Table 3. Key Outcomes by Funding, Disease, Randomization, and Location

Outcome	Studies, No. (%)							
	Funding		Disease		Type of study		Location	
	Industry exclusively	Other sources or combinations	COVID-19	Other	Randomized	Nonrandomized	US	Other or international
<b>Funding</b>								
Industry exclusively	NA	NA	24 (22.4) <sup>a</sup>	279 (43.4) <sup>a</sup>	241 (51.5)	62 (47.0)	31 (36.1) <sup>a</sup>	272 (52.9) <sup>a</sup>
Other sources or combinations	NA	NA	83 (77.6) <sup>a</sup>	214 (56.6) <sup>a</sup>	227 (48.5)	70 (53.0)	55 (63.9) <sup>a</sup>	242 (47.1) <sup>a</sup>
<b>Industry affiliation for any author</b>								
Yes	279 (92.1) <sup>a</sup>	75 (25.2) <sup>a</sup>	51 (47.7) <sup>b</sup>	303 (61.5) <sup>b</sup>	269 (57.5)	85 (64.4)	36 (41.9) <sup>a</sup>	318 (61.9) <sup>a</sup>
No	24 (7.9) <sup>a</sup>	222 (74.8) <sup>a</sup>	56 (52.3) <sup>b</sup>	190 (38.5) <sup>b</sup>	199 (42.5)	47 (35.6)	50 (58.1) <sup>a</sup>	196 (38.1) <sup>a</sup>
<b>Analysts' affiliation</b>								
Only by industry analysts	113 (37.3) <sup>a</sup>	12 (4.0) <sup>a</sup>	21 (19.6)	104 (21.1)	107 (22.9) <sup>b</sup>	18 (13.6) <sup>b</sup>	11 (12.8) <sup>b</sup>	114 (22.2) <sup>b</sup>
Other	190 (62.7) <sup>a</sup>	285 (96.0) <sup>a</sup>	86 (80.4)	389 (78.9)	361 (77.1) <sup>b</sup>	114 (86.4) <sup>b</sup>	75 (87.2) <sup>b</sup>	400 (77.8) <sup>b</sup>
<b>Access to data</b>								
Data are available to others	1 (0.3) <sup>a</sup>	15 (5.0) <sup>a</sup>	2 (1.9)	14 (2.8)	11 (2.4)	5 (3.8)	8 (9.3) <sup>a</sup>	8 (1.6) <sup>a</sup>
No access or other ways to access	302 (99.7) <sup>a</sup>	282 (95.0) <sup>a</sup>	105 (98.1)	479 (97.2)	457 (97.6)	127 (96.2)	78 (90.7) <sup>a</sup>	506 (98.4) <sup>a</sup>
<b>Full protocol availability</b>								
Yes	278 (91.8) <sup>a</sup>	214 (72.1) <sup>a</sup>	83 (77.6)	409 (83.0)	405 (86.5) <sup>a</sup>	87 (65.9) <sup>a</sup>	65 (75.6)	427 (83.1)
No	25 (8.2) <sup>a</sup>	83 (27.9) <sup>a</sup>	24 (22.4)	84 (17.0)	63 (13.5) <sup>a</sup>	45 (34.1) <sup>a</sup>	21 (24.4)	87 (16.9)
<b>Statistical analysis plan availability</b>								
Yes	262 (86.5) <sup>a</sup>	184 (62.0) <sup>a</sup>	69 (64.5)	377 (76.5)	373 (79.7) <sup>a</sup>	73 (55.3) <sup>a</sup>	56 (65.1) <sup>b</sup>	390 (75.9) <sup>b</sup>
No	41 (13.5) <sup>a</sup>	113 (38.0) <sup>a</sup>	38 (35.5)	116 (23.5)	95 (20.3) <sup>a</sup>	59 (44.7) <sup>a</sup>	30 (34.9) <sup>b</sup>	124 (24.1) <sup>b</sup>
<b>Results favoring sponsor (for industry-funded trials, n = 409)</b>								
Yes	279 (92.1) <sup>a</sup>	85 (80.2) <sup>a</sup>	38 (79.2) <sup>b</sup>	326 (90.3) <sup>b</sup>	274 (86.4) <sup>a</sup>	90 (97.8) <sup>a</sup>	44 (97.8) <sup>b</sup>	320 (87.9) <sup>b</sup>
No	24 (7.9) <sup>a</sup>	21 (19.8) <sup>a</sup>	10 (20.8) <sup>b</sup>	35 (9.7) <sup>b</sup>	43 (13.6) <sup>a</sup>	2 (2.2) <sup>a</sup>	1 (2.2) <sup>b</sup>	44 (12.1) <sup>b</sup>

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



Naudet et al, BMJ 2018

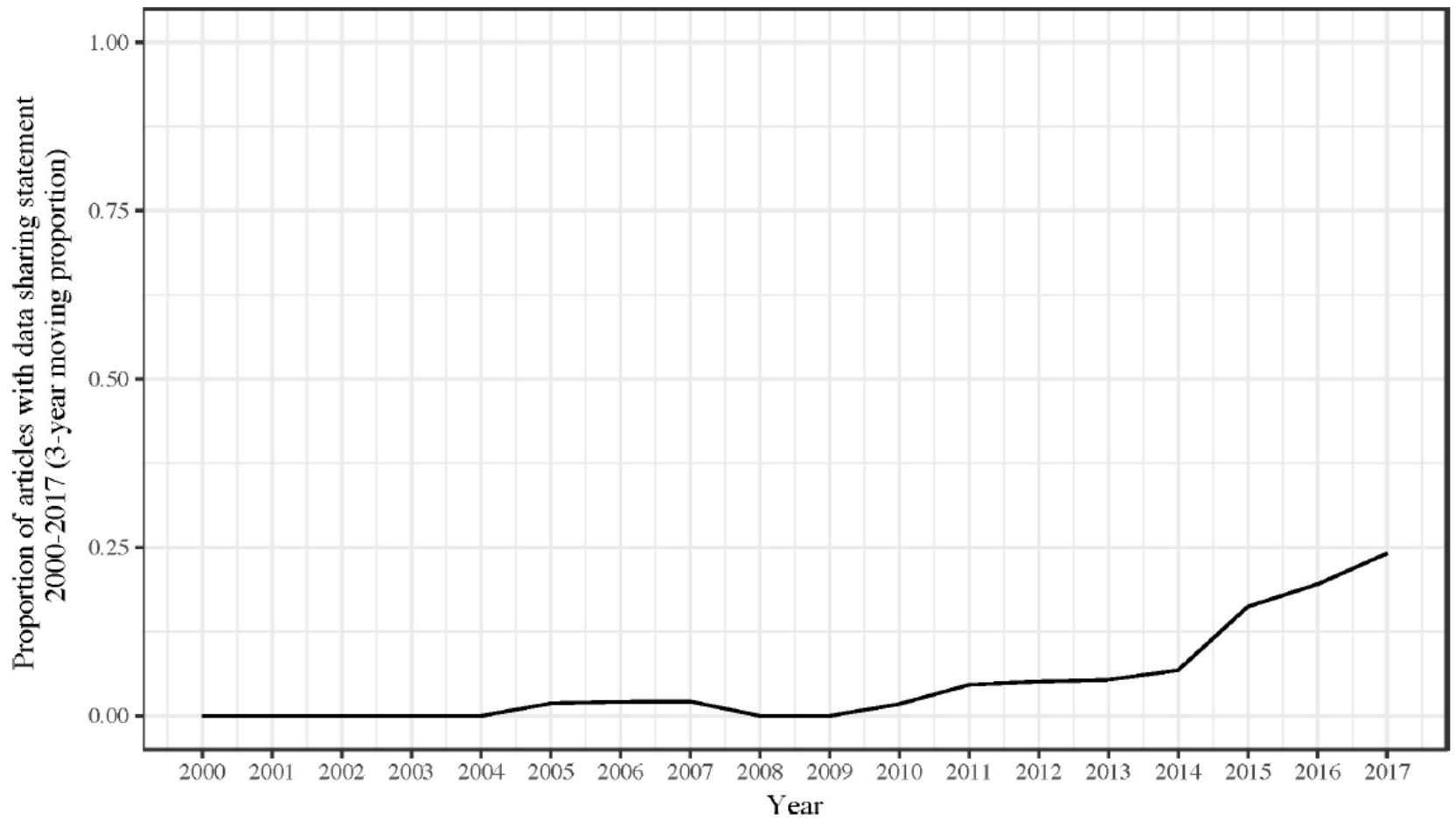
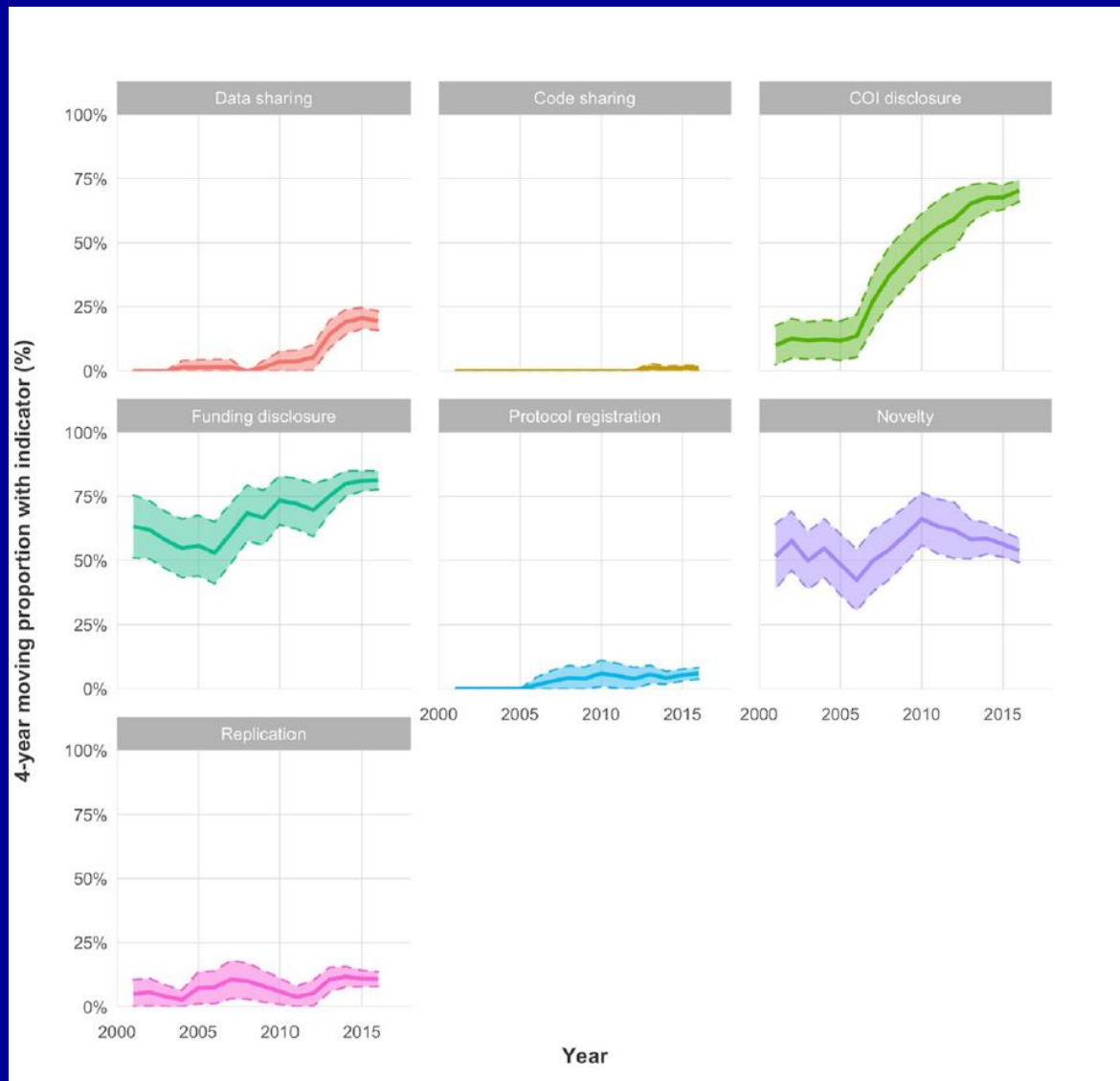


Fig 2. Proportion of articles with data sharing statement, 2000–2017 (3-year moving proportion). Underlying data for Fig 2 can be found at <https://osf.io/3ypdn/>.

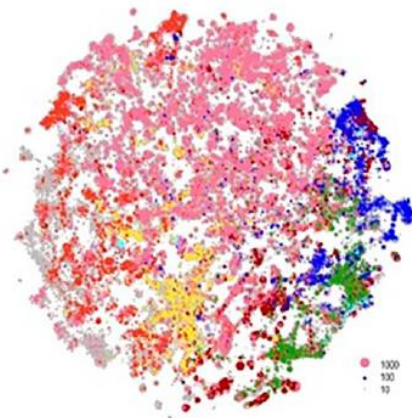
<https://doi.org/10.1371/journal.pbio.2006930.g002>

# Assessment of transparency indicators across the biomedical literature: How open is open?

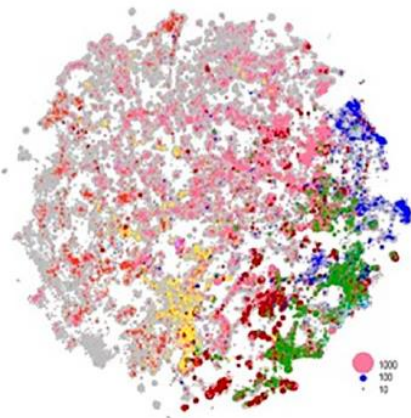
Stylianos Serghiou<sup>1,2</sup>, Despina G. Contopoulos-Ioannidis<sup>3</sup>, Kevin W. Boyack<sup>4</sup>,  
Nico Riedel<sup>5</sup>, Joshua D. Wallach<sup>6</sup>, John P. A. Ioannidis<sup>1,2,7,8,9\*</sup>



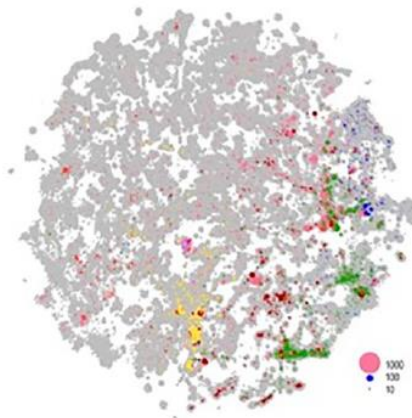
**Open Access**



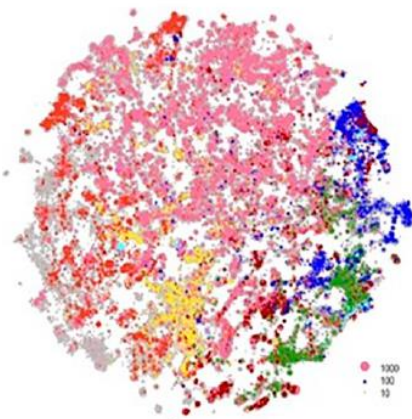
**Data sharing**



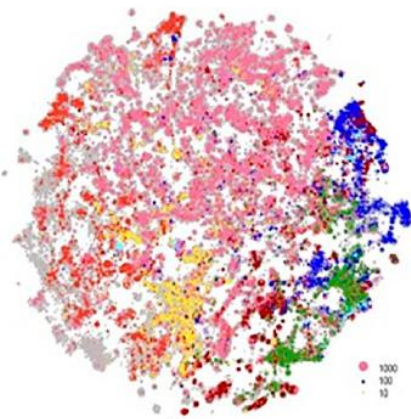
**Code sharing**



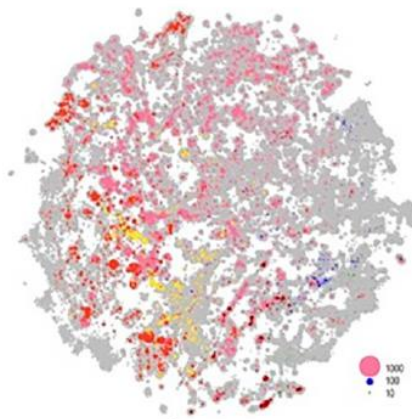
**COI disclosure**



**Funding disclosure**



**Protocol registration**



Physics

Computer

Chemistry

Engineering

Earth

Biology

Disease

Medicine

Brain

Health

Social

Humanities

REPRODUCIBILITY

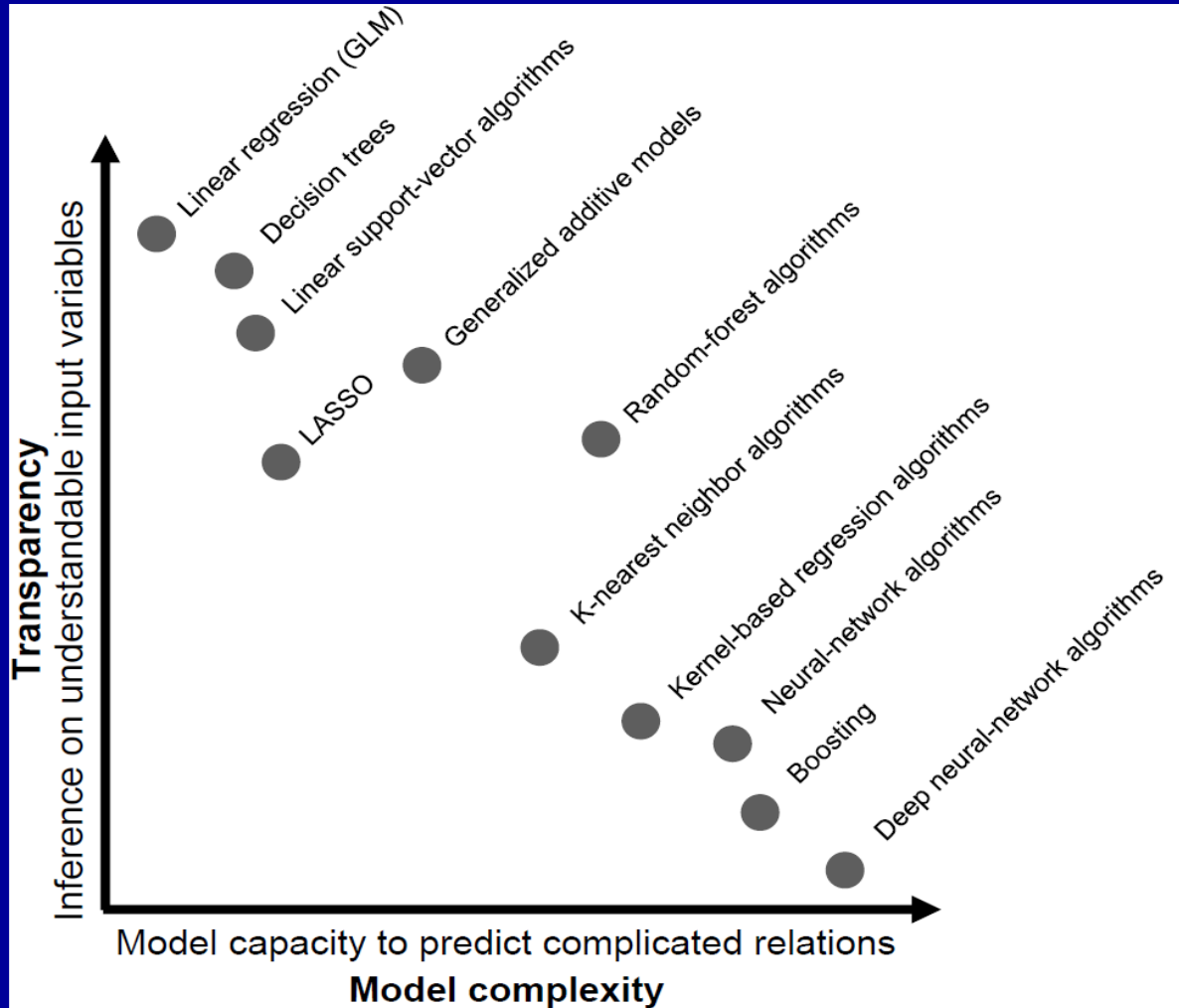
# Enhancing Reproducibility for Computational Methods

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*

Stodden et al. Science, 2016

# Transparency versus complexity




# Transparency and reproducibility in artificial intelligence

<https://doi.org/10.1038/s41586-020-2766-y>

Received: 1 February 2020

Accepted: 10 August 2020

Published online: 14 October 2020

 Check for updates

Benjamin Haibe-Kains<sup>1,2,3,4,5,6,7</sup>, George Alexandru Adam<sup>8,9</sup>, Ahmed Hosny<sup>6,7</sup>, Farnoosh Khodakarami<sup>10</sup>, Massive Analysis Quality Control (MAQC) Society Board of Directors<sup>\*</sup>, Levi Waldron<sup>6</sup>, Bo Wang<sup>2,3,5,9,10</sup>, Chris McIntosh<sup>2,5,7</sup>, Anna Goldenberg<sup>3,5,11,12</sup>, Anshul Kundaje<sup>13,14</sup>, Casey S. Greene<sup>15,16</sup>, Tamara Broderick<sup>17</sup>, Michael M. Hoffman<sup>1,2,3,5</sup>, Jeffrey T. Leek<sup>18</sup>, Keegan Korthauer<sup>19,20</sup>, Wolfgang Huber<sup>21</sup>, Alvis Brazma<sup>22</sup>, Joelle Pineau<sup>23,24</sup>, Robert Tibshirani<sup>25,26</sup>, Trevor Hastie<sup>25,26</sup>, John P. A. Ioannidis<sup>25,26,27,28,29</sup>, John Quackenbush<sup>30,31,32</sup> & Hugo J. W. L. Aerts<sup>6,7,32,34</sup>

ARISING FROM S. M. McKinney et al. *Nature* <https://doi.org/10.1038/s41586-019-1799-6> (2020)

**Table 1 | Essential hyperparameters for reproducing the study for each of the three models**

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

publication, McKinney et al.<sup>1</sup> did not disclose the settings for the augmentation pipeline; the transformations used are stochastic and can considerably affect model performance<sup>30</sup>. Details of the training pipeline were also missing. Without this key information, independent reproduction of the training pipeline is not possible.

Numerous frameworks and platforms exist to make artificial intelligence research more transparent and reproducible (Table 2). For the sharing of code, these include Bitbucket, GitHub and GitLab, among others. The many software dependencies of large-scale machine learning applications require appropriate control of the software environment, which can be achieved through package managers including

**Table 2 | Frameworks to share code, software dependencies and deep-learning models**

Resource	URL
<b>Code</b>	
BitBucket	<a href="https://bitbucket.org">https://bitbucket.org</a>
GitHub	<a href="https://github.com">https://github.com</a>
GitLab	<a href="https://about.gitlab.com">https://about.gitlab.com</a>
<b>Software dependencies</b>	
Conda	<a href="https://conda.io">https://conda.io</a>
Code Ocean	<a href="https://codeocean.com">https://codeocean.com</a>
Gigantum	<a href="https://gigantum.com">https://gigantum.com</a>
Laboratory	<a href="https://colab.research.google.com">https://colab.research.google.com</a>
<b>Deep-learning models</b>	
TensorFlow Hub	<a href="https://www.tensorflow.org/hub">https://www.tensorflow.org/hub</a>
ModelHub	<a href="http://modelhub.ai">http://modelhub.ai</a>
ModelDepot	<a href="https://modeldepot.io">https://modeldepot.io</a>
Model Zoo	<a href="https://modelzoo.co">https://modelzoo.co</a>
<b>Deep-learning frameworks</b>	
TensorFlow	<a href="https://www.tensorflow.org/">https://www.tensorflow.org/</a>
Caffe	<a href="https://caffe.berkeleyvision.org/">https://caffe.berkeleyvision.org/</a>
PyTorch	<a href="https://pytorch.org/">https://pytorch.org/</a>



# Large language models for science and medicine


Amalio Telenti<sup>1,2</sup> | Michael Auli<sup>3</sup> | Brian L. Hie<sup>3,4</sup> | Cyrus Maher<sup>2</sup> | Suchi Saria<sup>5</sup> |  
John P. A. Ioannidis<sup>6,7,8,9,10</sup> 

TABLE 1 Major limitations and challenges and potential mitigation and adoption solutions.

Limitations and challenges	Mitigation and adoption solutions
Poor performance	Need for increased awareness of the problems, higher human involvement in verification and validation of LLMs' output, anticipated improvements in LLM technology and training sets, enhanced transparency, judicious human involvement in assessment of transparency and in decision-making, performance of rigorous studies that assess the validity and overall performance of different ways to combine LLMs and human expertise or other interaction
Misinformation	Awareness of the problem, enhancing transparency on sources of information and their validity, consideration of improving the regulatory and legal tools
Inequalities and other societal impact	Consideration of open-source technology options and democratically controlled technology options, transparency regarding conflicts of interest, healthy scepticism towards potentially biased expertise (human, artificial intelligence or both), avoidance of monoculture thinking, tolerance to alternative views
Impact on scientific ecosystem	Realistic expectations and careful examination of gains and losses with different LLM technologies, anticipation of potential changes in workforce requirements, training and continued education and re-education, strengthening of rigorous research practices (in particular reproducibility and transparency), high-standards for the scientific rewards and incentive system that are aligned with rigorous research, optimizing and standardizing the rules for use and for characterization of misuse of LLMs
Ethics	Detailed consideration of the ethical challenges, creation of rigorous and relevant ethical, legal and regulatory framework that follows the pace of LLM evolution, blocking the toxic uses
Broader issues	Remain vigilant about opportunities and threats, perform rigorous research on LLMs and their applications to assess impact, utility and diverse repercussions



## International Congress on Peer Review and Scientific Publication

*Enhancing the quality and credibility of science*

[About](#) [9th Congress](#) [Past Congresses](#)

**The 10th International Congress on Peer Review and Scientific Publication will be held at the Swissôtel in Chicago, September 3-5, 2025**

Our aim is to encourage research into the quality and credibility of peer review and scientific publication, to establish the evidence base on

# Peer Review and Scientific Publication at a Crossroads

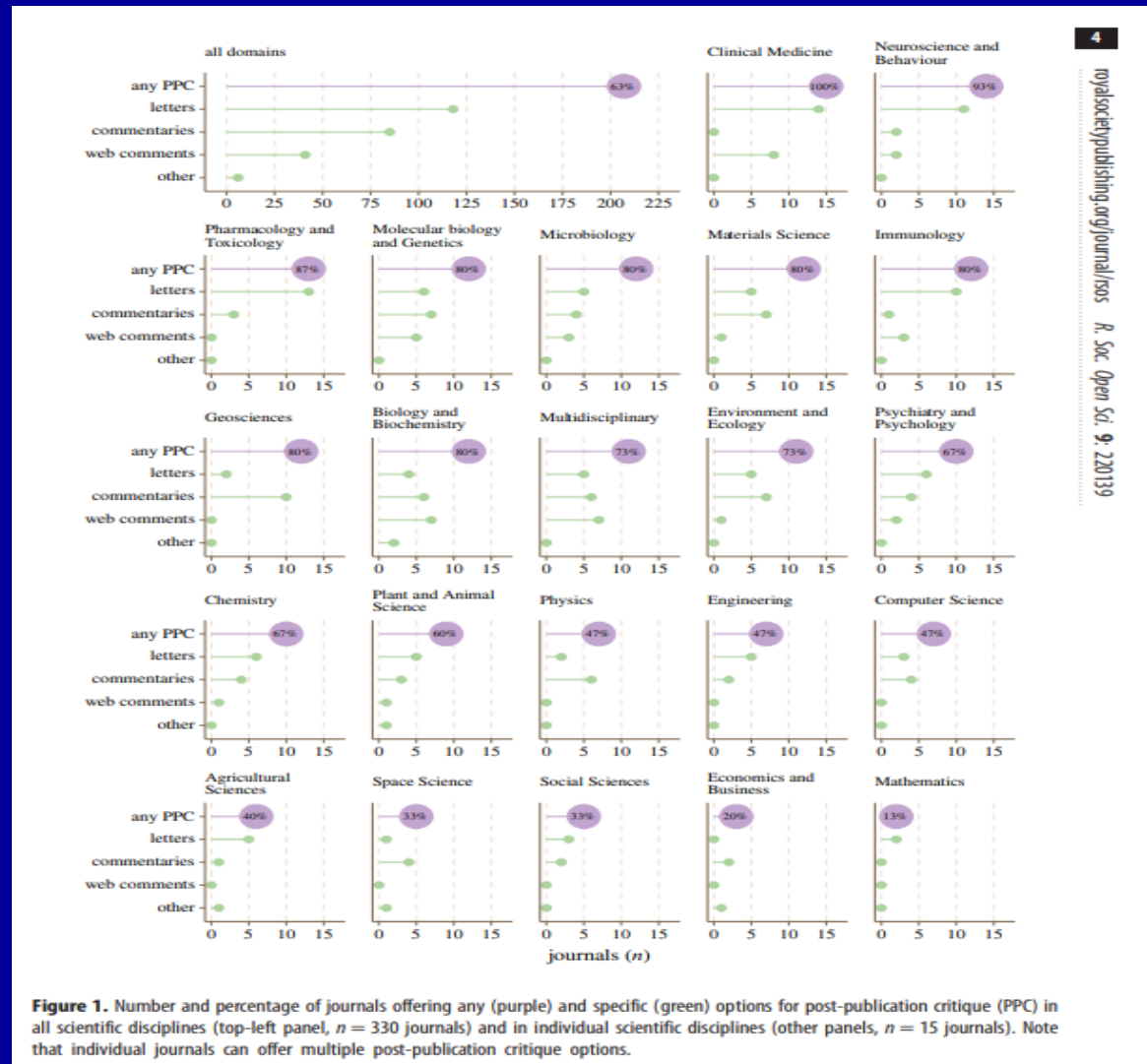
## Call for Research for the 10th International Congress on Peer Review and Scientific Publication

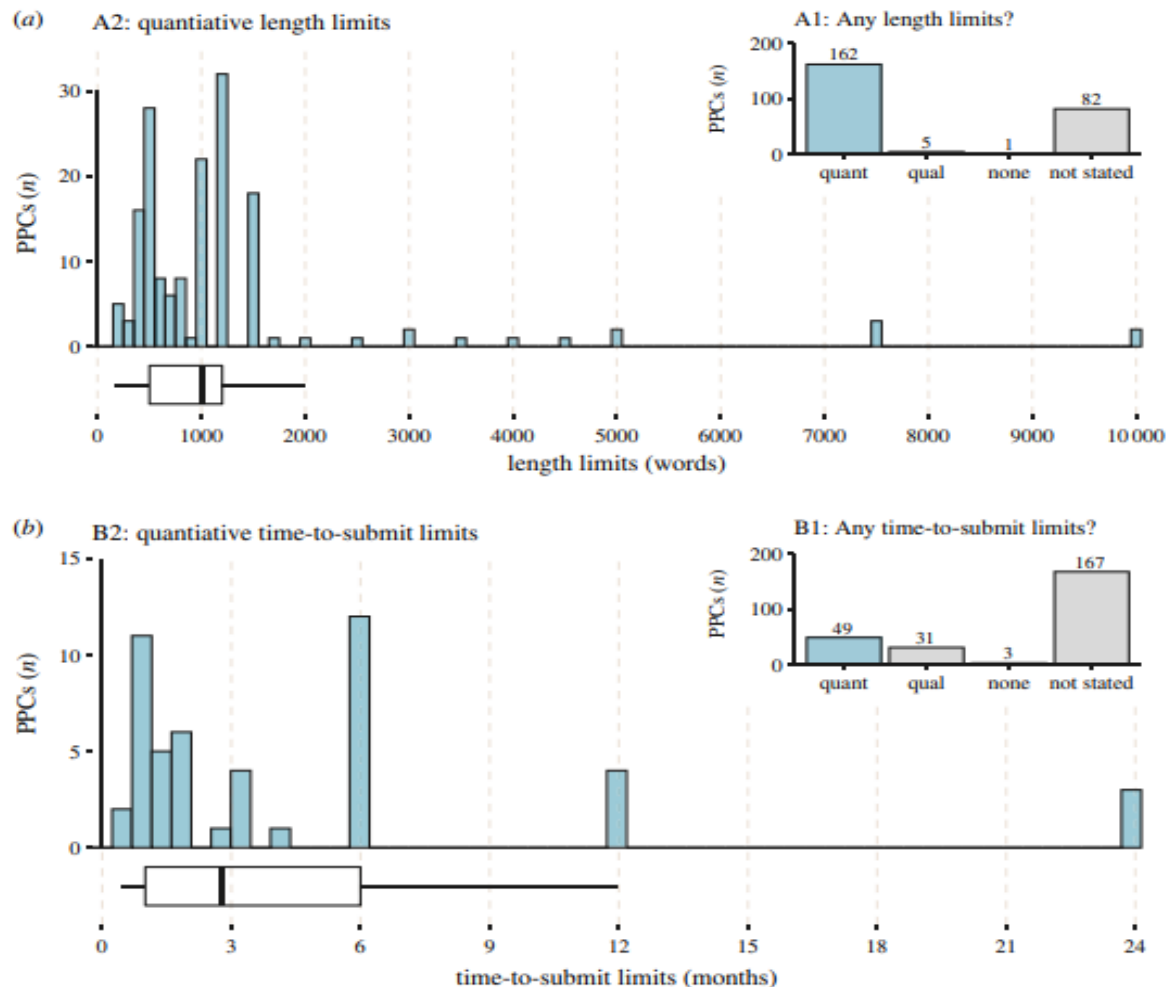
John P. A. Ioannidis, MD, DSc; Michael Berkwits, MD, MSCE; Annette Flanagan, RN, MA; Theodora Bloom, PhD

# Post-publication critique

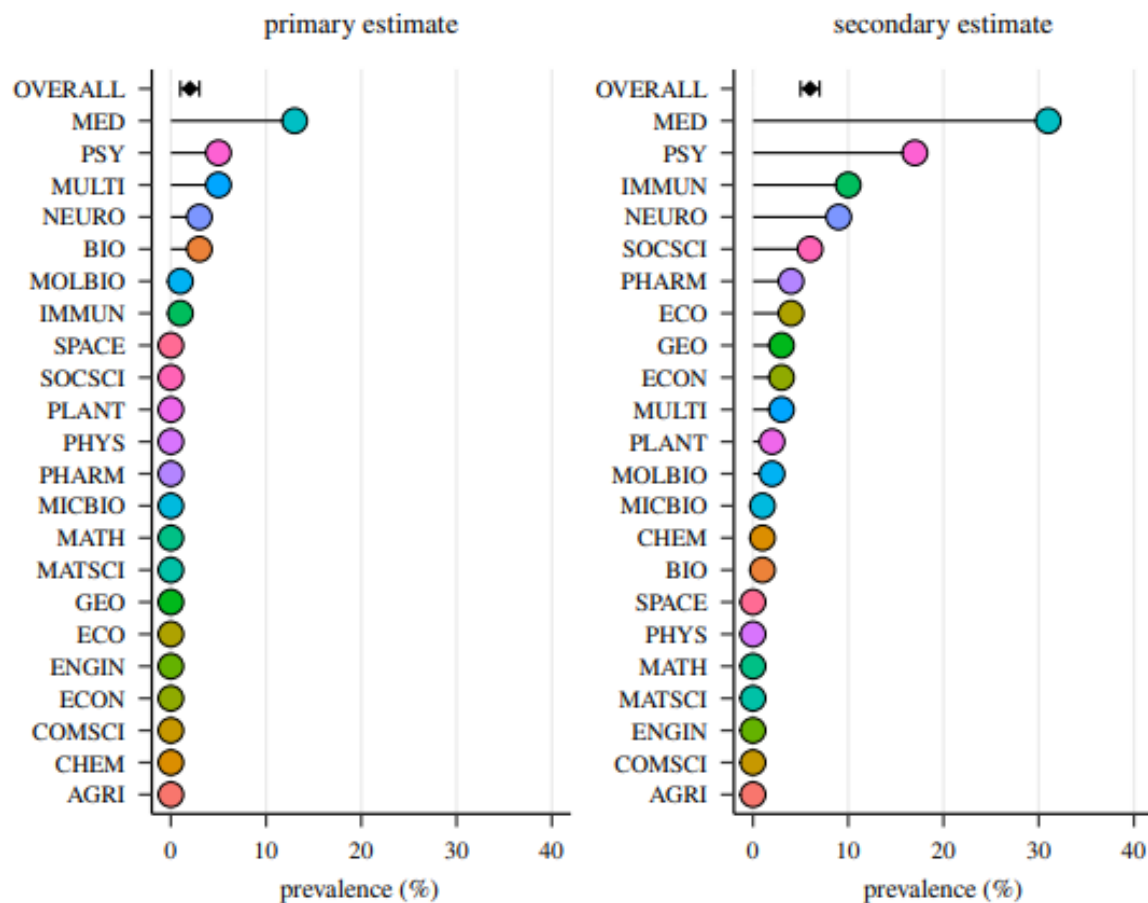
## Post-publication critique at top-ranked journals across scientific disciplines: a cross-sectional assessment of policies and practice

Tom E. Hardwicke<sup>1</sup>, Robert T. Thibault<sup>2,4</sup>,  
 Jessica E. Kosie<sup>5</sup>, Loukia Tzavella<sup>6</sup>, Theiss Bendixen<sup>7</sup>,  
 Sarah A. Handcock<sup>8</sup>, Vivian E. Köneke<sup>9</sup> and  
 John P. A. Ioannidis<sup>2,3,10</sup>





**Figure 2.** Limits imposed by journals on post-publication critique (PPC) in terms of (a) length and (b) time-to-submit since publication of the target article. A1 and B1 show the number of post-publication critique options for which the journal did not state if there was a limit (Not stated), explicitly stated there was not a limit (None), stated a qualitative limit (Qual) or stated a quantitative limit (Quant). Quantitative limits are displayed in A2 and B2 as a histogram and boxplot with the dark line representing the median, lower and upper hinges representing the 25th and 75th percentiles, and upper and lower whiskers representing the  $\pm 1.5$  interquartile range.



**Figure 3.** Primary (a) and secondary (b) prevalence estimates for post-publication critique in all journals overall ( $N = 330$  journals; black diamond, error bars represent 95% confidence intervals) and then in descending order by each scientific discipline ( $n = 15$  journals; coloured circles). Discipline abbreviations: Agricultural Sciences (AGRI), Biology and Biochemistry (BIO), Chemistry (CHEM), Clinical Medicine (MED), Computer Science (COMSCI), Economics and Business (ECON), Engineering (ENGIN), Environment and Ecology (ECO), Geosciences (GEO), Immunology (IMMUN), Materials Science (MATSCI), Mathematics (MATH), Microbiology (MICBIO), Molecular Biology and Genetics (MOLBIO), Multidisciplinary (MULTI), Neuroscience and Behaviour (NEURO), Pharmacology and Toxicology (PHARM), Physics (PHYS), Plant and Animal Science (PLANT), Psychiatry and Psychology (PSY), Social Sciences (SOCSCI), Space Science (SPACE).

# 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 $\geq 1$ 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 $\geq 1$ 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

# Assessing scientists for hiring, promotion, and tenure

**David Moher**<sup>1,2\*</sup>, **Florian Naudet**<sup>2,3</sup>, **Ioana A. Cristea**<sup>2,4</sup>, **Frank Miedema**<sup>5</sup>, **John P. A. Ioannidis**<sup>2,6,7,8,9</sup>, **Steven N. Goodman**<sup>2,6,7</sup>

**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](mailto: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.

# Academic criteria for promotion and tenure in biomedical sciences faculties: cross sectional analysis of international sample of universities

Danielle B Rice,<sup>1,2</sup> Hana Raffoul,<sup>2,3</sup> John P A Ioannidis,<sup>4,5,6,7</sup> David Moher<sup>8,9</sup>

## ABSTRACT

### OBJECTIVE

To determine the presence of a set of pre-specified traditional and non-traditional criteria used to assess scientists for promotion and tenure in faculties of biomedical sciences among universities worldwide.

### DESIGN

Cross sectional study.

### SETTING

International sample of universities.

### PARTICIPANTS

170 randomly selected universities from the Leiden ranking of world universities list.

### MAIN OUTCOME MEASURE

Presence of five traditional (for example, number of publications) and seven non-traditional (for example, data sharing) criteria in guidelines for assessing assistant professors, associate professors, and professors and the granting of tenure in institutions with biomedical faculties.

### RESULTS

A total of 146 institutions had faculties of biomedical sciences, and 92 had eligible guidelines available for review. Traditional criteria of peer reviewed publications, authorship order, journal impact factor, grant funding, and national or international reputation were mentioned in 95% (n=87), 37% (34), 28% (26), 67% (62), and 48% (44) of the guidelines, respectively. Conversely, among non-traditional

criteria, only citations (any mention in 26%; n=24) and accommodations for employment leave (37%; 34) were relatively commonly mentioned. Mention of alternative metrics for sharing research (3%; n=3) and data sharing (1%; 1) was rare, and three criteria (publishing in open access mediums, registering research, and adhering to reporting guidelines) were not found in any guidelines reviewed. Among guidelines for assessing promotion to full professor, traditional criteria were more commonly reported than non-traditional criteria (traditional criteria 54.2%, non-traditional items 9.5%; mean difference 44.8%, 95% confidence interval 39.6% to 50.0%; P=0.001). Notable differences were observed across continents in whether guidelines were accessible (Australia 100% (6/6), North America 97% (28/29), Europe 50% (27/54), Asia 58% (29/50), South America 17% (1/6)), with more subtle differences in the use of specific criteria.

### CONCLUSIONS

This study shows that the evaluation of scientists emphasises traditional criteria as opposed to non-traditional criteria. This may reinforce research practices that are known to be problematic while insufficiently supporting the conduct of better quality research and open science. Institutions should consider incentivising non-traditional criteria.

### STUDY REGISTRATION

Open Science Framework ([https://osf.io/26ucp/?view\\_only=b80d2bc7416543639f577c1b8f756e44](https://osf.io/26ucp/?view_only=b80d2bc7416543639f577c1b8f756e44)).



# Concluding comments

- The discussion surrounding reproducibility and how to improve it has been intense
- Reproducibility indicators are surrogates; what matters in research, science and its positive impact is more complex
- There are new stakeholders and new ways of publishing science that may change fundamental notions about what the scientific record is
- Progress on reproducibility is in the eye of the beholder