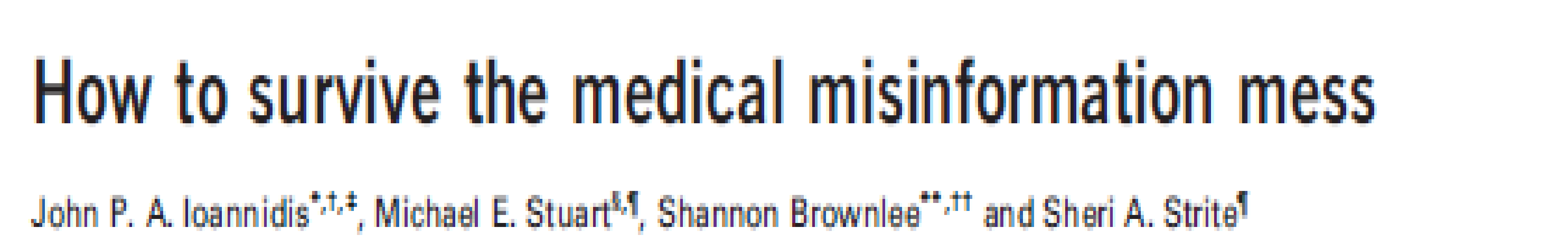
False and not useful clinical research: how can we increase its value?

TAYPOMENION, 10/2017

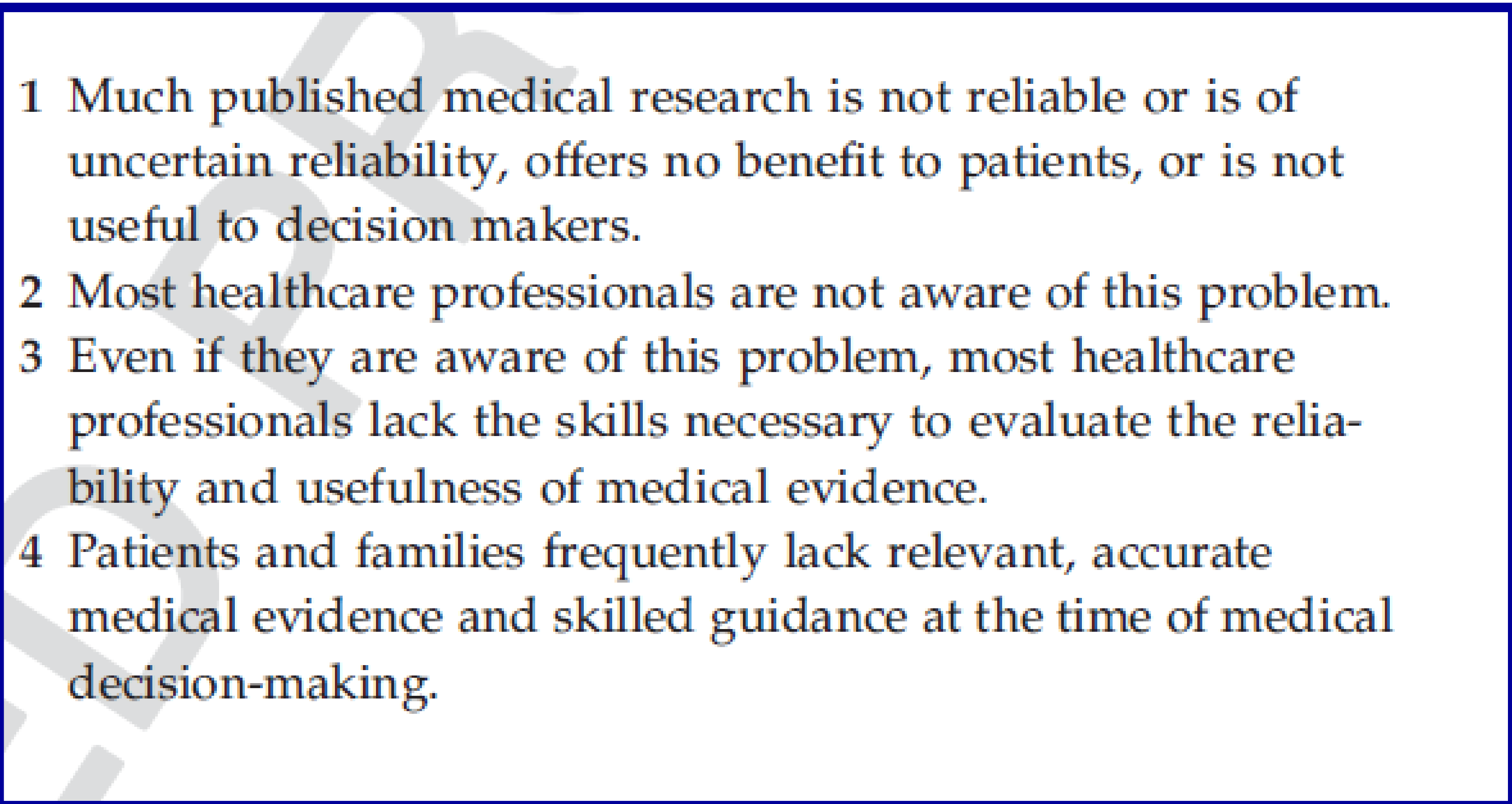
John P.A. Ioannidis, MD, DSc

C.F. Rehnborg Chair in Disease Prevention
Professor of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics
Stanford University
Co-Director, Meta-Research Innovation Center at Stanford (METRICS)



How to survive the medical misinformation mess

John P. A. Ioannidis[*,†,‡], Michael E. Stuart[§,¶], Shannon Brownlee[**,††] and Sheri A. Strite[¶]



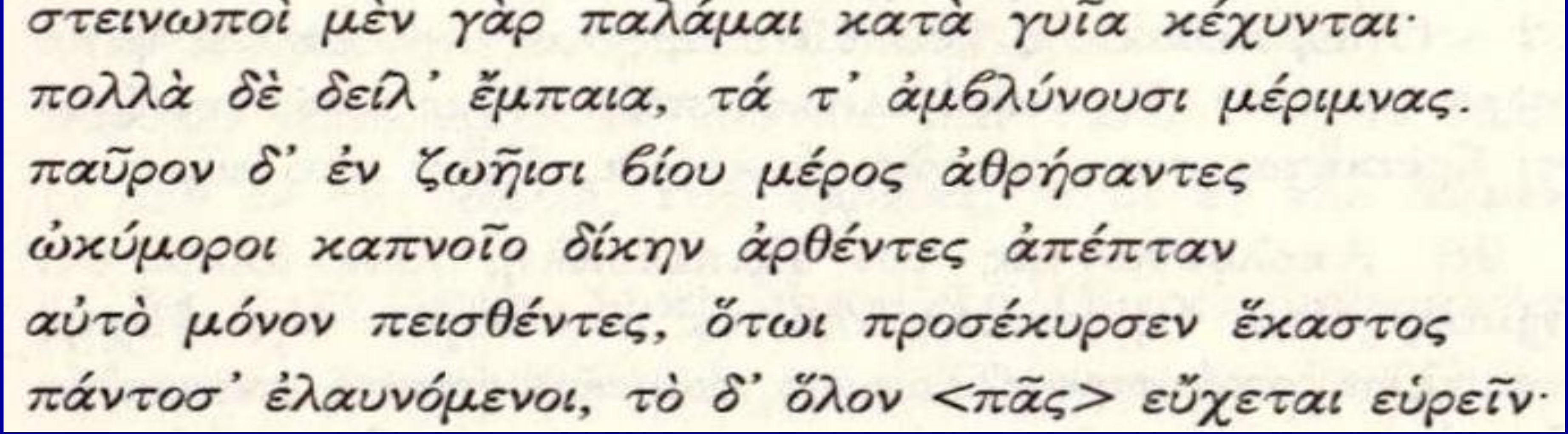
1 Much published medical research is not reliable or is of uncertain reliability, offers no benefit to patients, or is not useful to decision makers.

2 Most healthcare professionals are not aware of this problem.

3 Even if they are aware of this problem, most healthcare professionals lack the skills necessary to evaluate the reliability and usefulness of medical evidence.

4 Patients and families frequently lack relevant, accurate medical evidence and skilled guidance at the time of medical decision-making.

Eur J Clin Invest, 2017



στεινωποὶ μὲν γὰρ παλάμαι κατὰ γυῖα κέχυνται·
πολλὰ δὲ δείλ' ἔμπαια, τά τ' ἀμβλύνουσι μέριμνας.
παῦρον δ' ἐν ζωῆισι βίου μέρος ἀθρήσαντες
ὠκύμοροι καπνοῖο δίκην ἀρθέντες ἀπέπταν
αὐτὸ μόνον πεισθέντες, ὅτωι προσέκυρσεν ἕκαστος
πάντοσ' ἐλαυνόμενοι, τὸ δ' ὅλον <πᾶς> εὔχεται εὑρεῖν·

ΕΜΠΕΔΟΚΛΗΣ

For scant and scattered are the means of acquiring evidence. And many sad happenings intervene that blunt the edge of careful reasoning. After gathering only a small portion of life that is not life, swift to meet their fate, they get dispersed like smoke, persuaded only of whatever bias each one of them chanced upon while being tossed around here and there, boasting in vain to have found the whole.

Friedrich Hölderlin

Der Tod des Empedokles

Ein Trauerspiel in fünf Akten

HOFENBERG

How good is the quality of the clinical evidence?

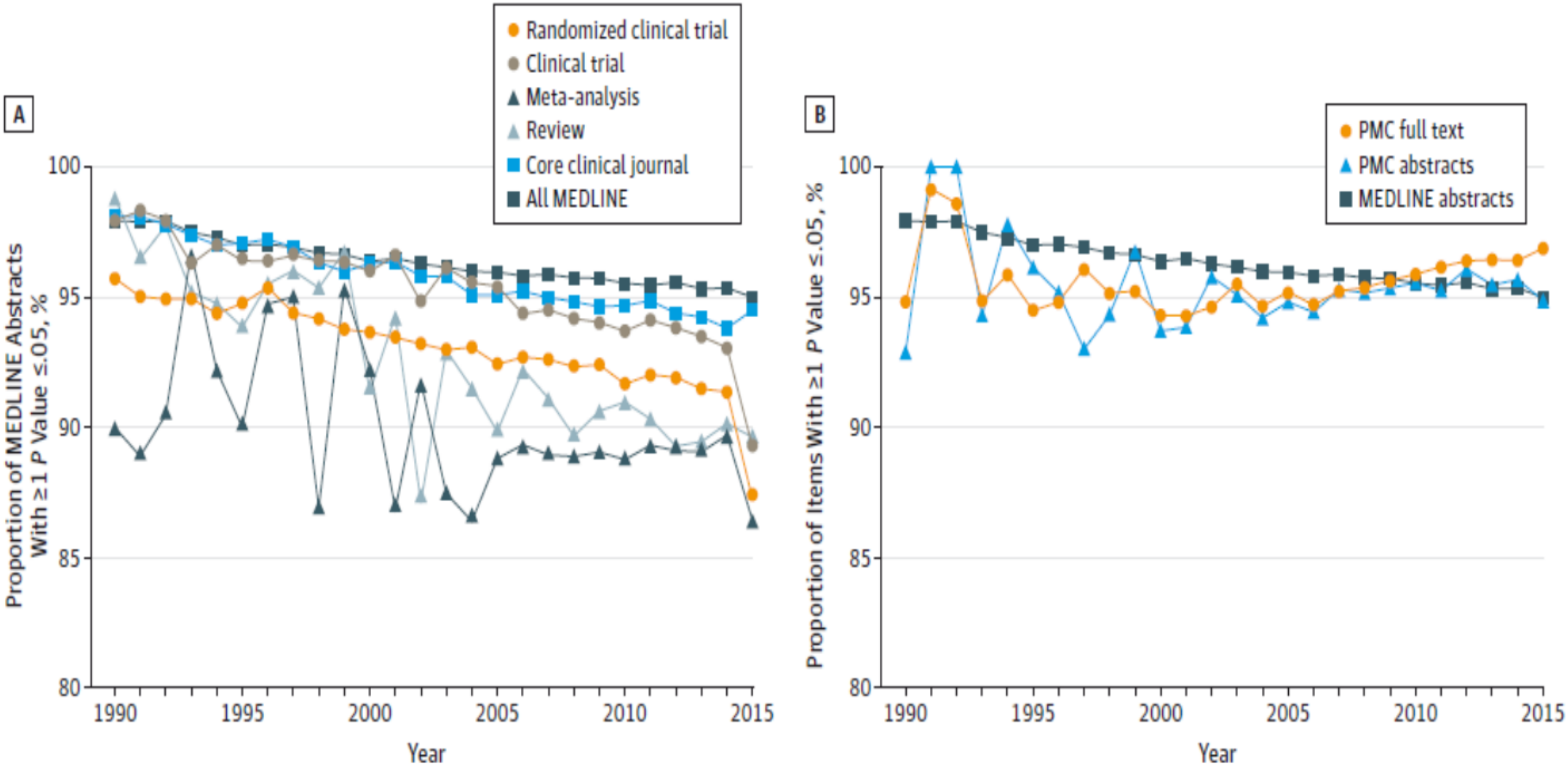
- All 1394 systematic reviews published on the Cochrane Database of Systematic Reviews from January 2013 to June, 2014.
- GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) summary of findings performed in 608 (43.6%).
- Quality of the evidence for the first listed primary outcome: 13.5% high, 30.8% moderate, 31.7% low, 24% very low level.
- Even when all outcomes listed were considered, only 19.1% had at least one outcome with high quality of evidence.
- Of the reviews with high quality of evidence, only 25 had both significant results and a favorable interpretation of the intervention.

Fleming et al, J Clin Epidemiol 2016

Significance of the evidence?

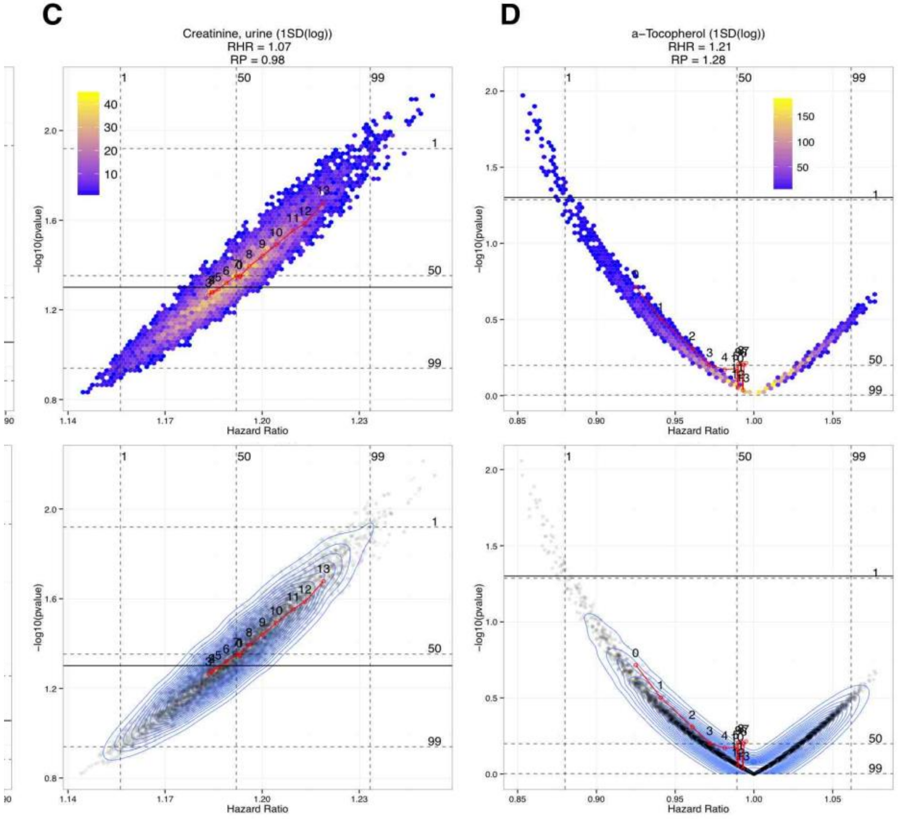
- Almost all scientific papers claim that they have found (statistically and/or conceptually) significant results
- Among abstracts with P-values in Medline (1990-2015), 96% report statistically significant results

Statistical significance has become a boring nuisance: 96% of the biomedical literature claims significant results



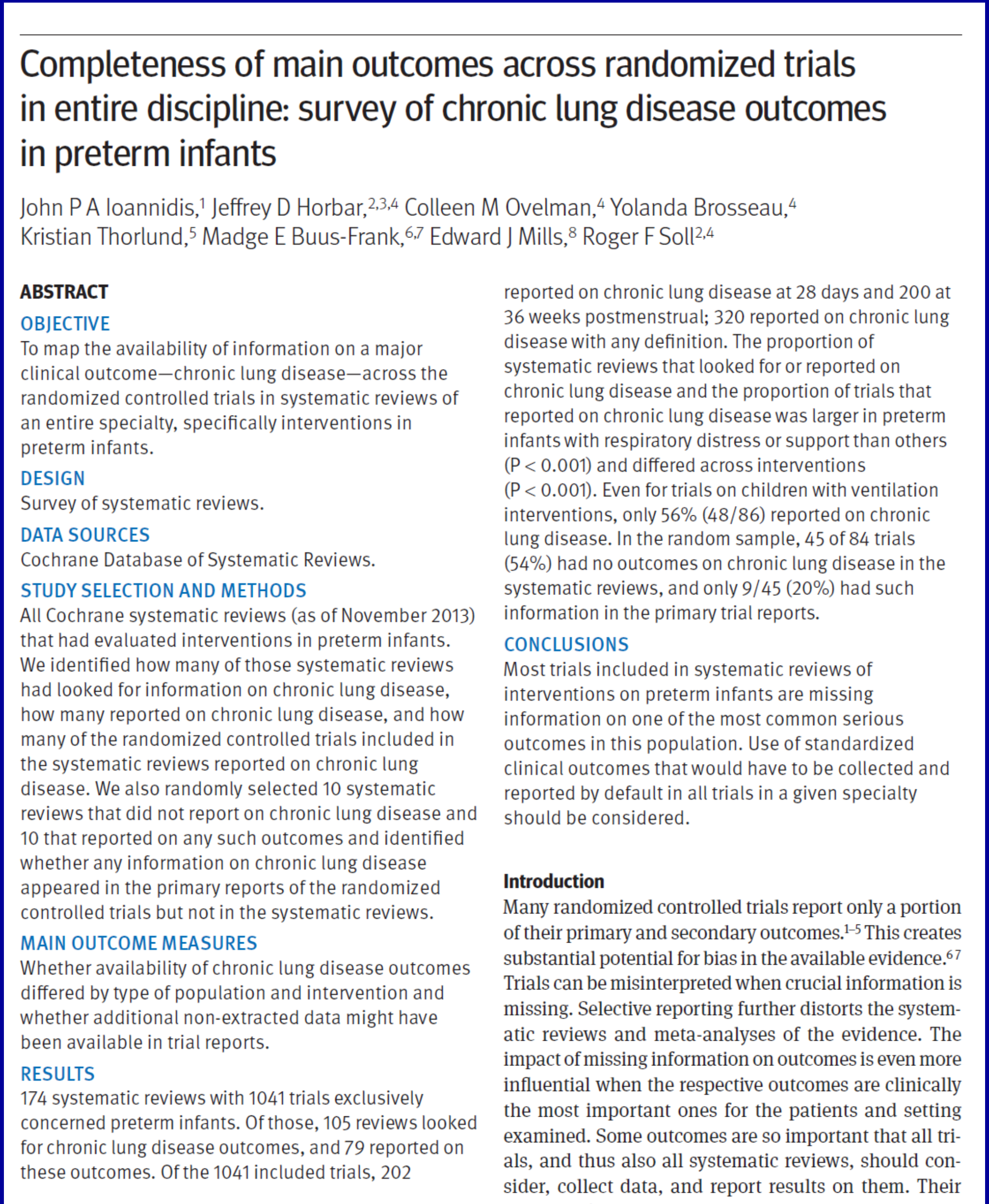
Chavalarias, Wallach, Li, Ioannidis, JAMA 2016

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



Patel, Burford, Ioannidis. JCE 2015

Patient-relevant outcomes are understudied



Completeness of main outcomes across randomized trials in entire discipline: survey of chronic lung disease outcomes in preterm infants

John P A Ioannidis,[1] Jeffrey D Horbar,[2,3,4] Colleen M Ovelman,[4] Yolanda Brosseau,[4]
Kristian Thorlund,[5] Madge E Buus-Frank,[6,7] Edward J Mills,[8] Roger F Soll[2,4]

ABSTRACT

OBJECTIVE

To map the availability of information on a major clinical outcome—chronic lung disease—across the randomized controlled trials in systematic reviews of an entire specialty, specifically interventions in preterm infants.

DESIGN

Survey of systematic reviews.

DATA SOURCES

Cochrane Database of Systematic Reviews.

STUDY SELECTION AND METHODS

All Cochrane systematic reviews (as of November 2013) that had evaluated interventions in preterm infants. We identified how many of those systematic reviews had looked for information on chronic lung disease, how many reported on chronic lung disease, and how many of the randomized controlled trials included in the systematic reviews reported on chronic lung disease. We also randomly selected 10 systematic reviews that did not report on chronic lung disease and 10 that reported on any such outcomes and identified whether any information on chronic lung disease appeared in the primary reports of the randomized controlled trials but not in the systematic reviews.

MAIN OUTCOME MEASURES

Whether availability of chronic lung disease outcomes differed by type of population and intervention and whether additional non-extracted data might have been available in trial reports.

RESULTS

174 systematic reviews with 1041 trials exclusively concerned preterm infants. Of those, 105 reviews looked for chronic lung disease outcomes, and 79 reported on these outcomes. Of the 1041 included trials, 202 reported on chronic lung disease at 28 days and 200 at 36 weeks postmenstrual; 320 reported on chronic lung disease with any definition. The proportion of systematic reviews that looked for or reported on chronic lung disease and the proportion of trials that reported on chronic lung disease was larger in preterm infants with respiratory distress or support than others ($P < 0.001$) and differed across interventions ($P < 0.001$). Even for trials on children with ventilation interventions, only 56% (48/86) reported on chronic lung disease. In the random sample, 45 of 84 trials (54%) had no outcomes on chronic lung disease in the systematic reviews, and only 9/45 (20%) had such information in the primary trial reports.

CONCLUSIONS

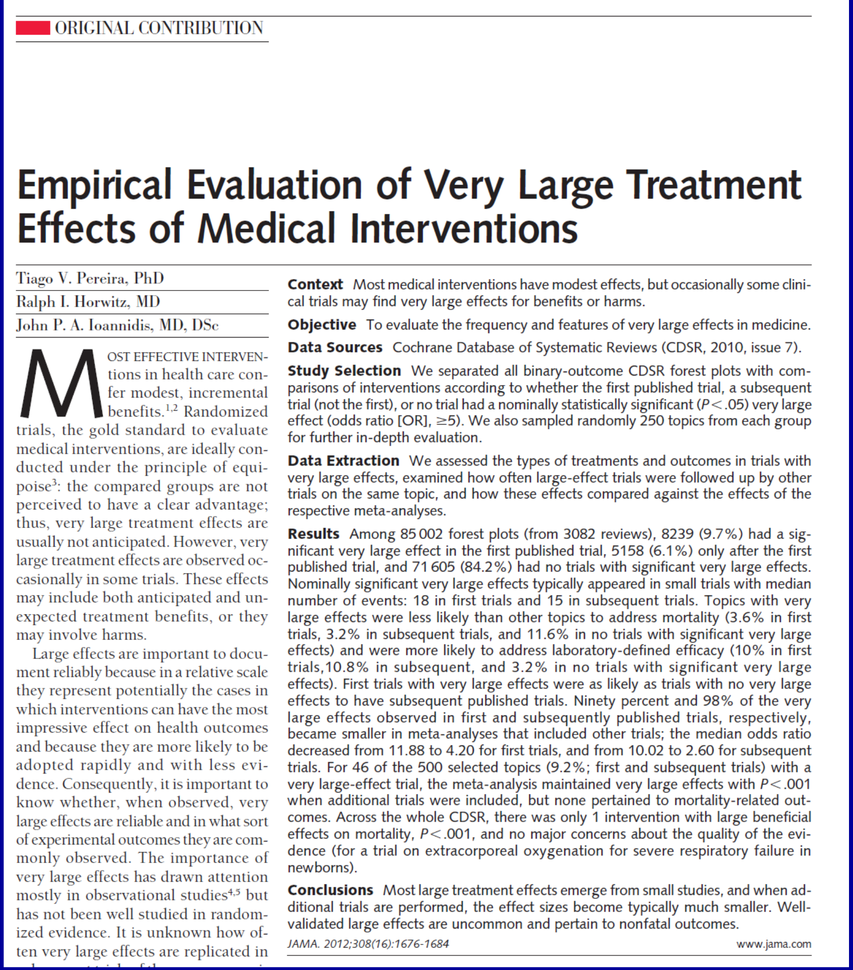
Most trials included in systematic reviews of interventions on preterm infants are missing information on one of the most common serious outcomes in this population. Use of standardized clinical outcomes that would have to be collected and reported by default in all trials in a given specialty should be considered.

Introduction

Many randomized controlled trials report only a portion of their primary and secondary outcomes.[1-5] This creates substantial potential for bias in the available evidence.[6,7] Trials can be misinterpreted when crucial information is missing. Selective reporting further distorts the systematic reviews and meta-analyses of the evidence. The impact of missing information on outcomes is even more influential when the respective outcomes are clinically the most important ones for the patients and setting examined. Some outcomes are so important that all trials, and thus also all systematic reviews, should consider, collect data, and report results on them. Their

Chronic lung disease in preterm infants reported in only 320/1041 trials

Many treatment effects seem to be large, especially in small, early trials, but few survive scrutiny



ORIGINAL CONTRIBUTION

Empirical Evaluation of Very Large Treatment Effects of Medical Interventions

Tiago V. Pereira, PhD
Ralph I. Horwitz, MD
John P. A. Ioannidis, MD, DSc

Context Most medical interventions have modest effects, but occasionally some clinical trials may find very large effects for benefits or harms.

Objective To evaluate the frequency and features of very large effects in medicine.

Data Sources Cochrane Database of Systematic Reviews (CDSR, 2010, issue 7).

Study Selection We separated all binary-outcome CDSR forest plots with comparisons of interventions according to whether the first published trial, a subsequent trial (not the first), or no trial had a nominally statistically significant ($P<.05$) very large effect (odds ratio [OR], ≥5). We also sampled randomly 250 topics from each group for further in-depth evaluation.

Data Extraction We assessed the types of treatments and outcomes in trials with very large effects, examined how often large-effect trials were followed up by other trials on the same topic, and how these effects compared against the effects of the respective meta-analyses.

Results Among 85 002 forest plots (from 3082 reviews), 8239 (9.7%) had a significant very large effect in the first published trial, 5158 (6.1%) only after the first published trial, and 71 605 (84.2%) had no trials with significant very large effects. Nominally significant very large effects typically appeared in small trials with median number of events: 18 in first trials and 15 in subsequent trials. Topics with very large effects were less likely than other topics to address mortality (3.6% in first trials, 3.2% in subsequent trials, and 11.6% in no trials with significant very large effects) and were more likely to address laboratory-defined efficacy (10% in first trials,10.8% in subsequent, and 3.2% in no trials with significant very large effects). First trials with very large effects were as likely as trials with no very large effects to have subsequent published trials. Ninety percent and 98% of the very large effects observed in first and subsequently published trials, respectively, became smaller in meta-analyses that included other trials; the median odds ratio decreased from 11.88 to 4.20 for first trials, and from 10.02 to 2.60 for subsequent trials. For 46 of the 500 selected topics (9.2%; first and subsequent trials) with a very large-effect trial, the meta-analysis maintained very large effects with $P<.001$ when additional trials were included, but none pertained to mortality-related outcomes. Across the whole CDSR, there was only 1 intervention with large beneficial effects on mortality, $P<.001$, and no major concerns about the quality of the evidence (for a trial on extracorporeal oxygenation for severe respiratory failure in newborns).

Conclusions Most large treatment effects emerge from small studies, and when additional trials are performed, the effect sizes become typically much smaller. Well-validated large effects are uncommon and pertain to nonfatal outcomes.

JAMA. 2012;308(16):1676-1684 www.jama.com

MOST EFFECTIVE INTERVENTIONS in health care confer modest, incremental benefits.[1,2] Randomized trials, the gold standard to evaluate medical interventions, are ideally conducted under the principle of equipoise[3]: the compared groups are not perceived to have a clear advantage; thus, very large treatment effects are usually not anticipated. However, very large treatment effects are observed occasionally in some trials. These effects may include both anticipated and unexpected treatment benefits, or they may involve harms.

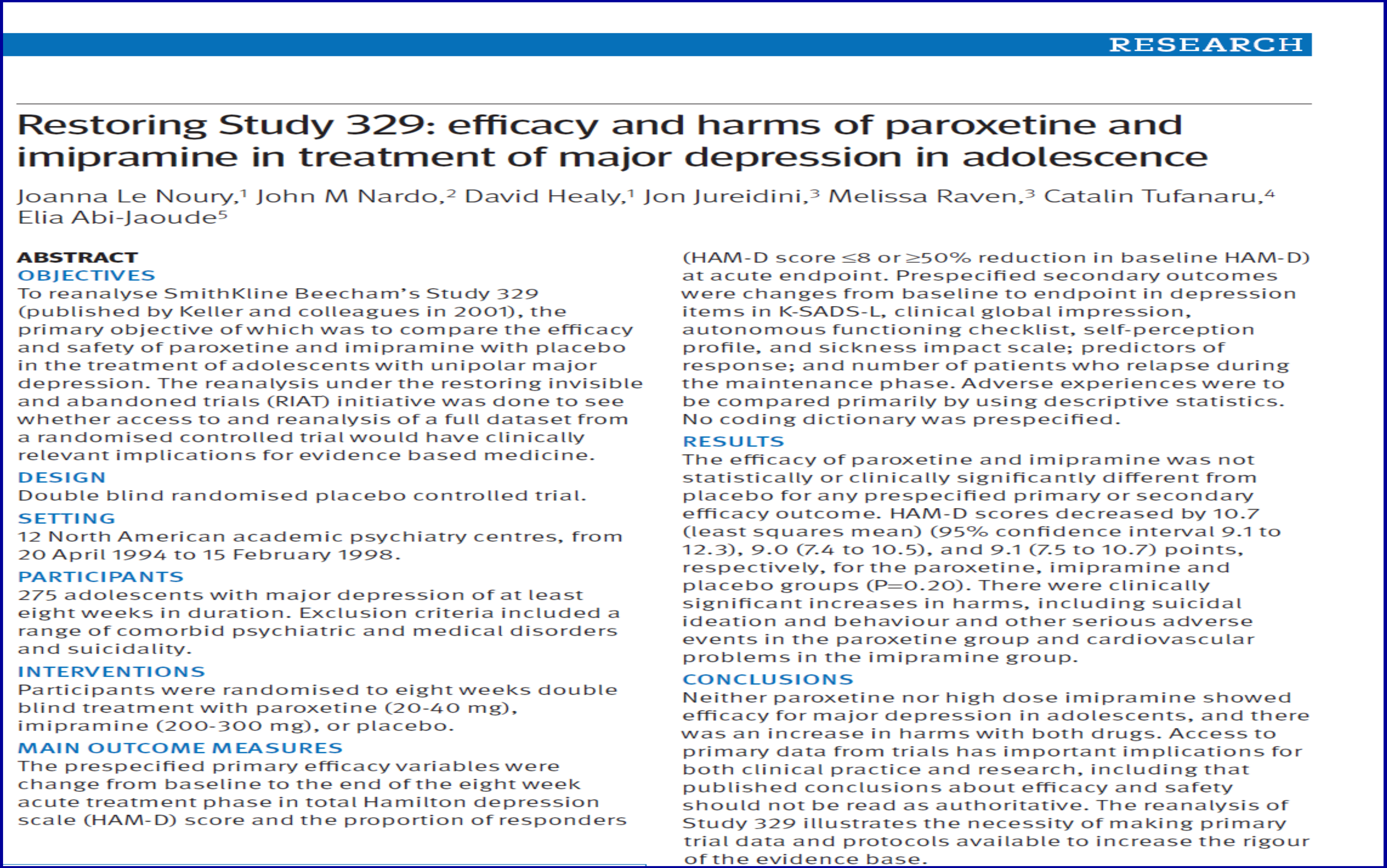
Large effects are important to document reliably because in a relative scale they represent potentially the cases in which interventions can have the most impressive effect on health outcomes and because they are more likely to be adopted rapidly and with less evidence. Consequently, it is important to know whether, when observed, very large effects are reliable and in what sort of experimental outcomes they are commonly observed. The importance of very large effects has drawn attention mostly in observational studies[4,5] but has not been well studied in randomized evidence. It is unknown how often very large effects are replicated in

Some types of clinical trials almost always favor the sponsor:

- Among trials published in 2011, 55/57 of non-inferiority trials with head to head comparisons sponsored by the industry demonstrated non-inferiority
- Success rate > 96%

Flacco et al. JCE 2015

Re-analysis: can we trust the data?



RESEARCH

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

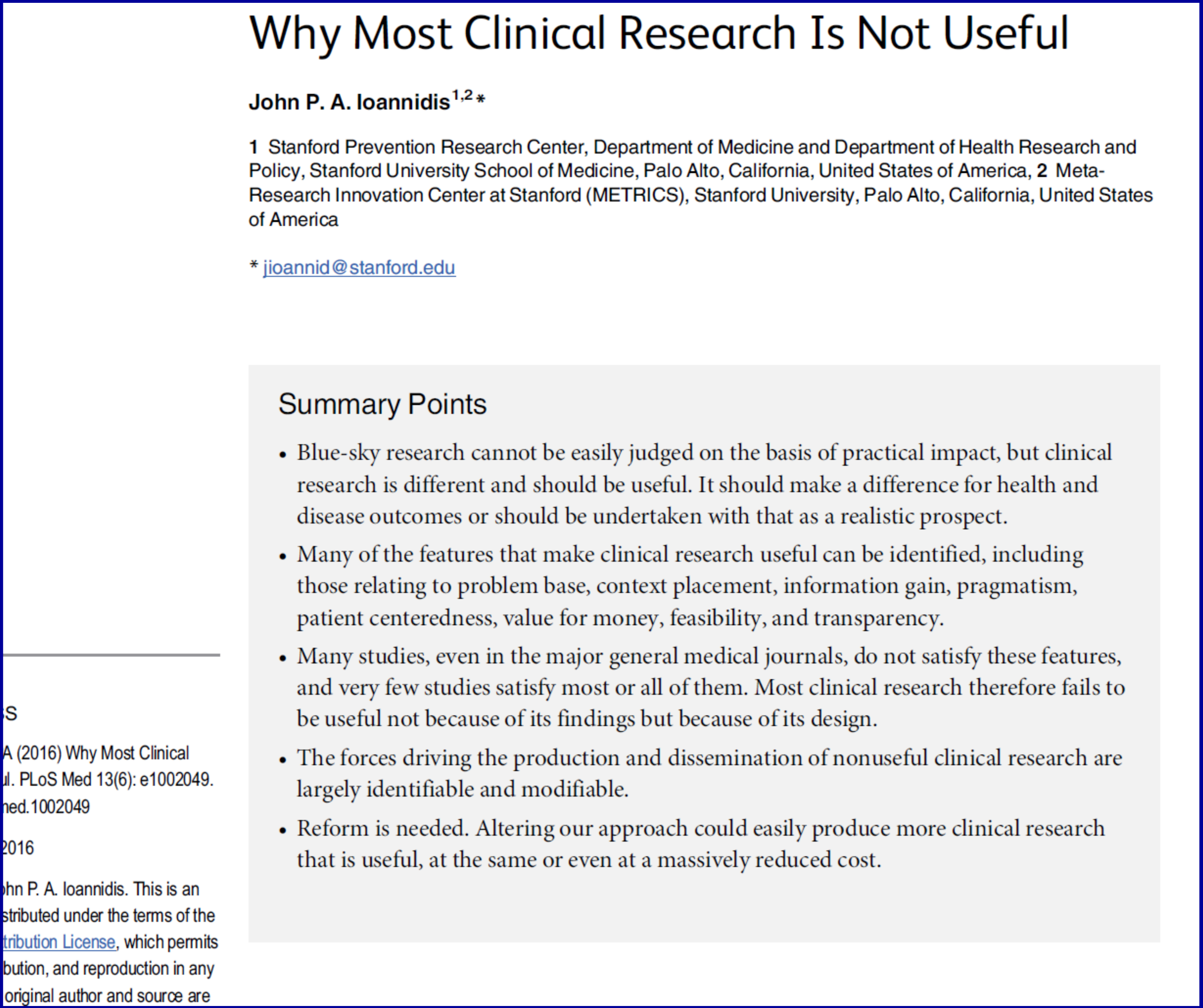
The prespecified primary efficacy variables were change from baseline to the end of the eight week acute treatment phase in total Hamilton depression scale (HAM-D) score and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Prespecified secondary outcomes 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.



Why Most Clinical Research Is Not Useful

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

1 Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, Palo Alto, California, United States of America, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, California, United States of America

* jioannid@stanford.edu

Summary Points

- Blue-sky research cannot be easily judged on the basis of practical impact, but clinical research is different and should be useful. It should make a difference for health and disease outcomes or should be undertaken with that as a realistic prospect.
- Many of the features that make clinical research useful can be identified, including those relating to problem base, context placement, information gain, pragmatism, patient centeredness, value for money, feasibility, and transparency.
- Many studies, even in the major general medical journals, do not satisfy these features, and very few studies satisfy most or all of them. Most clinical research therefore fails to be useful not because of its findings but because of its design.
- The forces driving the production and dissemination of nonuseful clinical research are largely identifiable and modifiable.
- Reform is needed. Altering our approach could easily produce more clinical research that is useful, at the same or even at a massively reduced cost.

s

A (2016) Why Most Clinical
l. PLoS Med 13(6): e1002049.
ned.1002049

2016

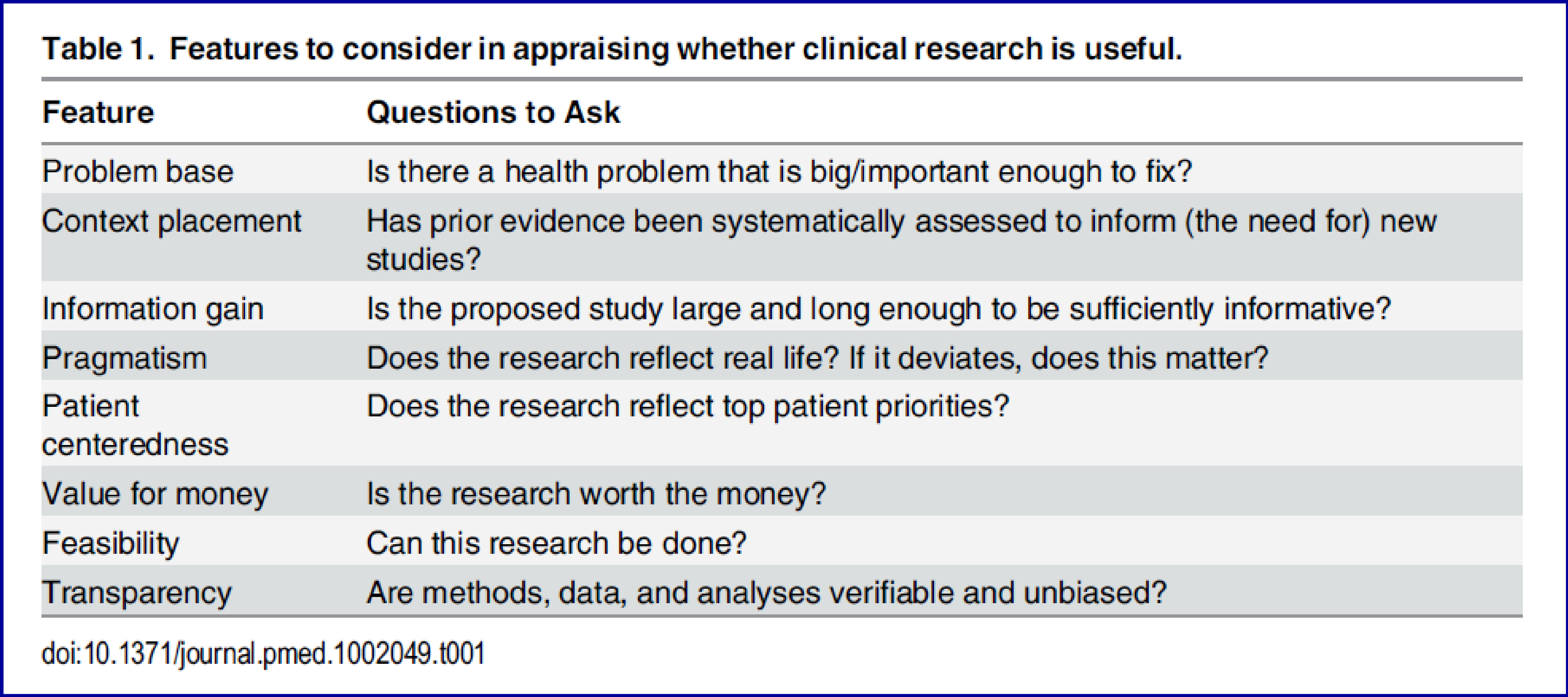


Table 1. Features to consider in appraising whether clinical research is useful.

Feature	Questions to Ask
Problem base	Is there a health problem that is big/important enough to fix?
Context placement	Has prior evidence been systematically assessed to inform (the need for) new studies?
Information gain	Is the proposed study large and long enough to be sufficiently informative?
Pragmatism	Does the research reflect real life? If it deviates, does this matter?
Patient centeredness	Does the research reflect top patient priorities?
Value for money	Is the research worth the money?
Feasibility	Can this research be done?
Transparency	Are methods, data, and analyses verifiable and unbiased?

doi:10.1371/journal.pmed.1002049.t001

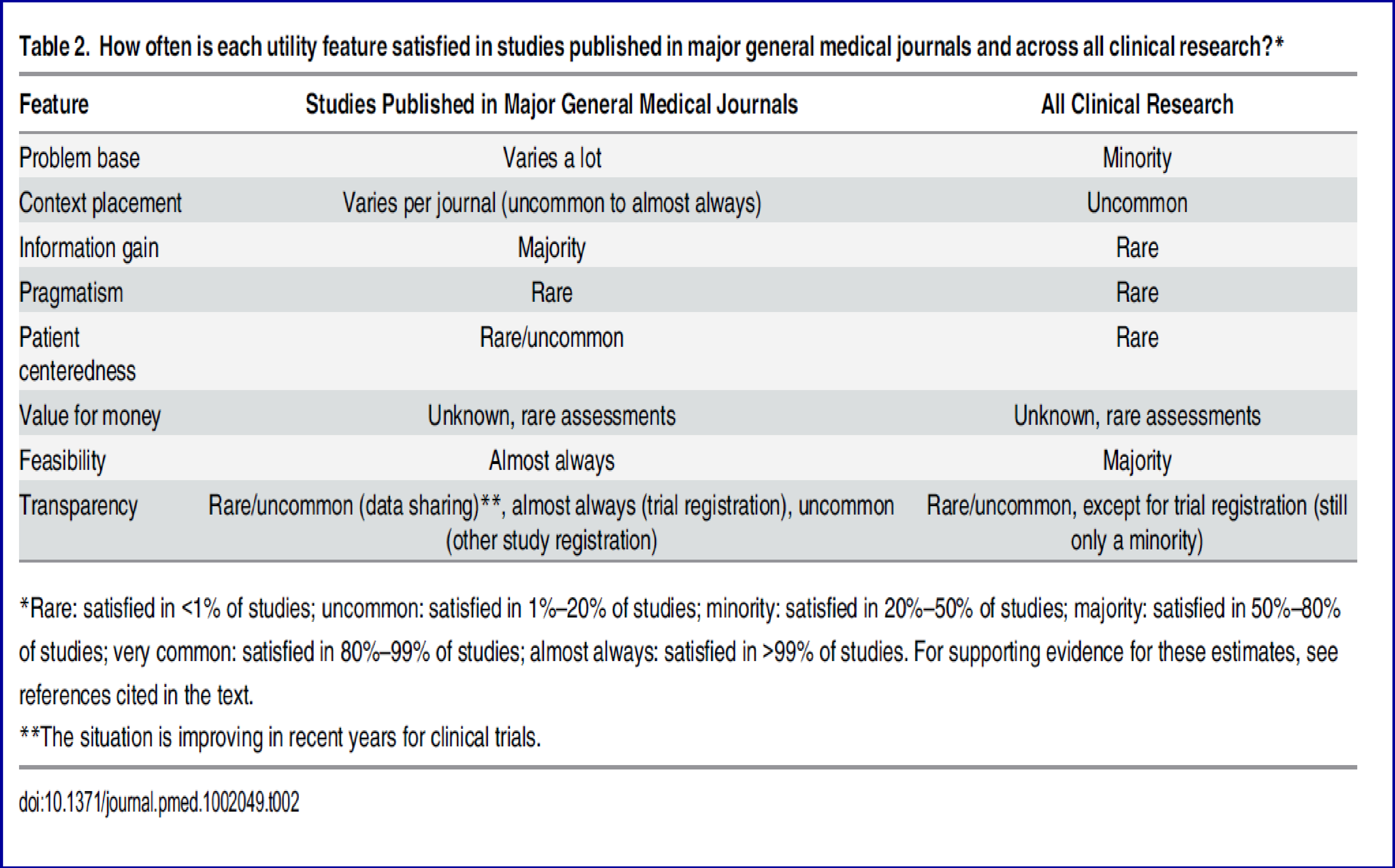


Table 2. How often is each utility feature satisfied in studies published in major general medical journals and across all clinical research?*

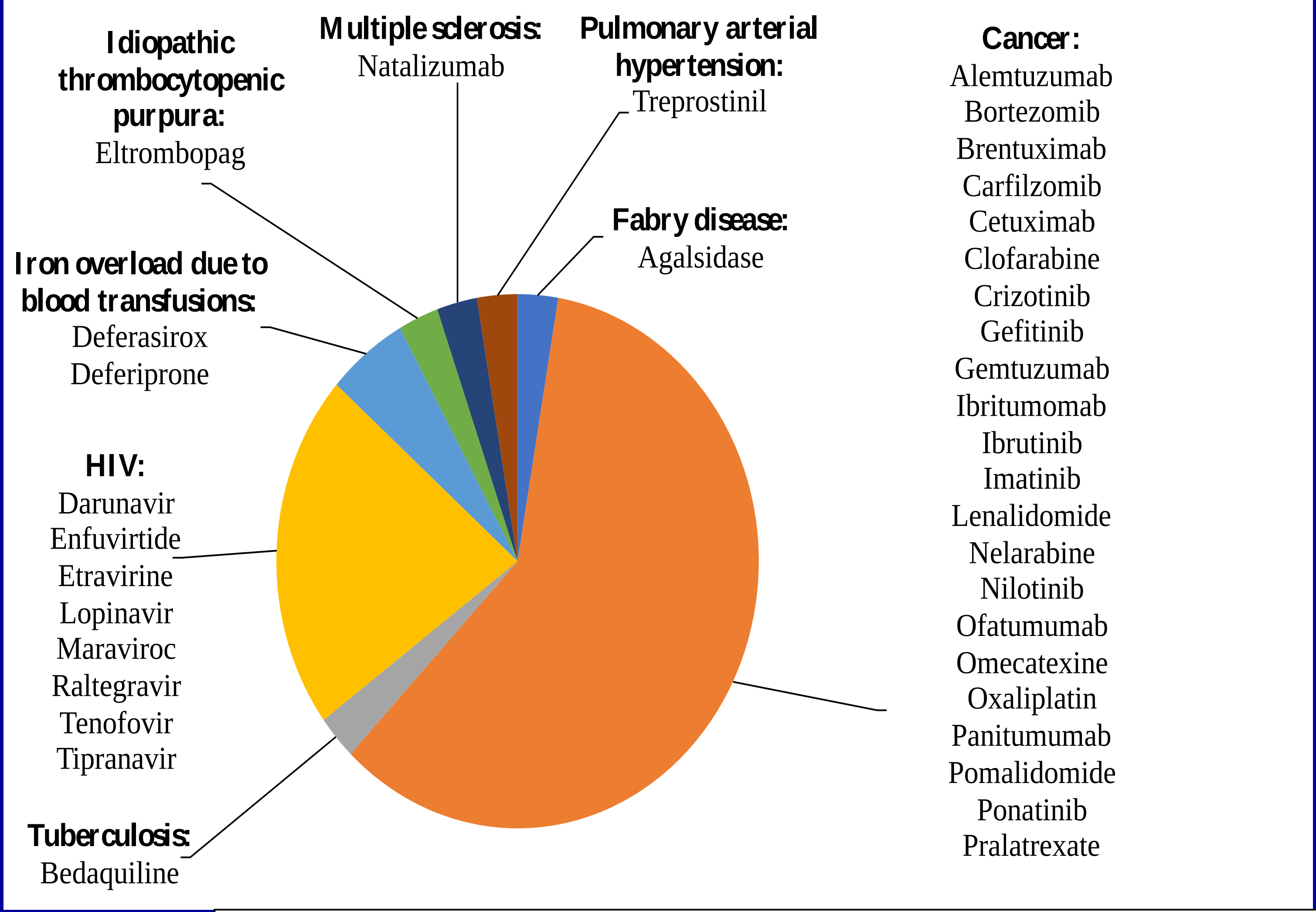
Feature	Studies Published in Major General Medical Journals	All Clinical Research
Problem base	Varies a lot	Minority
Context placement	Varies per journal (uncommon to almost always)	Uncommon
Information gain	Majority	Rare
Pragmatism	Rare	Rare
Patient centeredness	Rare/uncommon	Rare
Value for money	Unknown, rare assessments	Unknown, rare assessments
Feasibility	Almost always	Majority
Transparency	Rare/uncommon (data sharing)**, almost always (trial registration), uncommon (other study registration)	Rare/uncommon, except for trial registration (still only a minority)

*Rare: satisfied in <1% of studies; uncommon: satisfied in 1%–20% of studies; minority: satisfied in 20%–50% of studies; majority: satisfied in 50%–80% of studies; very common: satisfied in 80%–99% of studies; almost always: satisfied in >99% of studies. For supporting evidence for these estimates, see references cited in the text.

**The situation is improving in recent years for clinical trials.

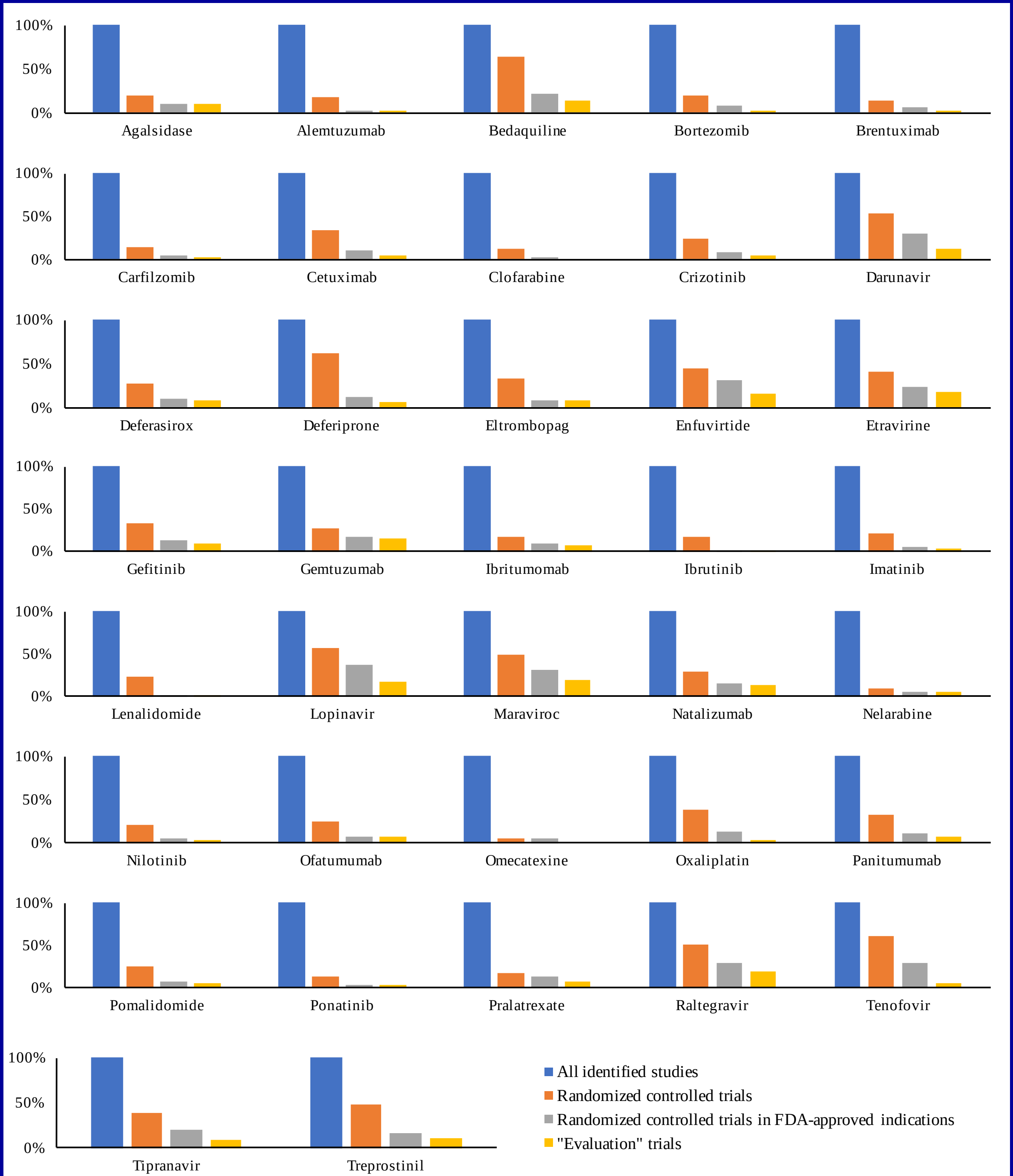
doi:10.1371/journal.pmed.1002049.t002

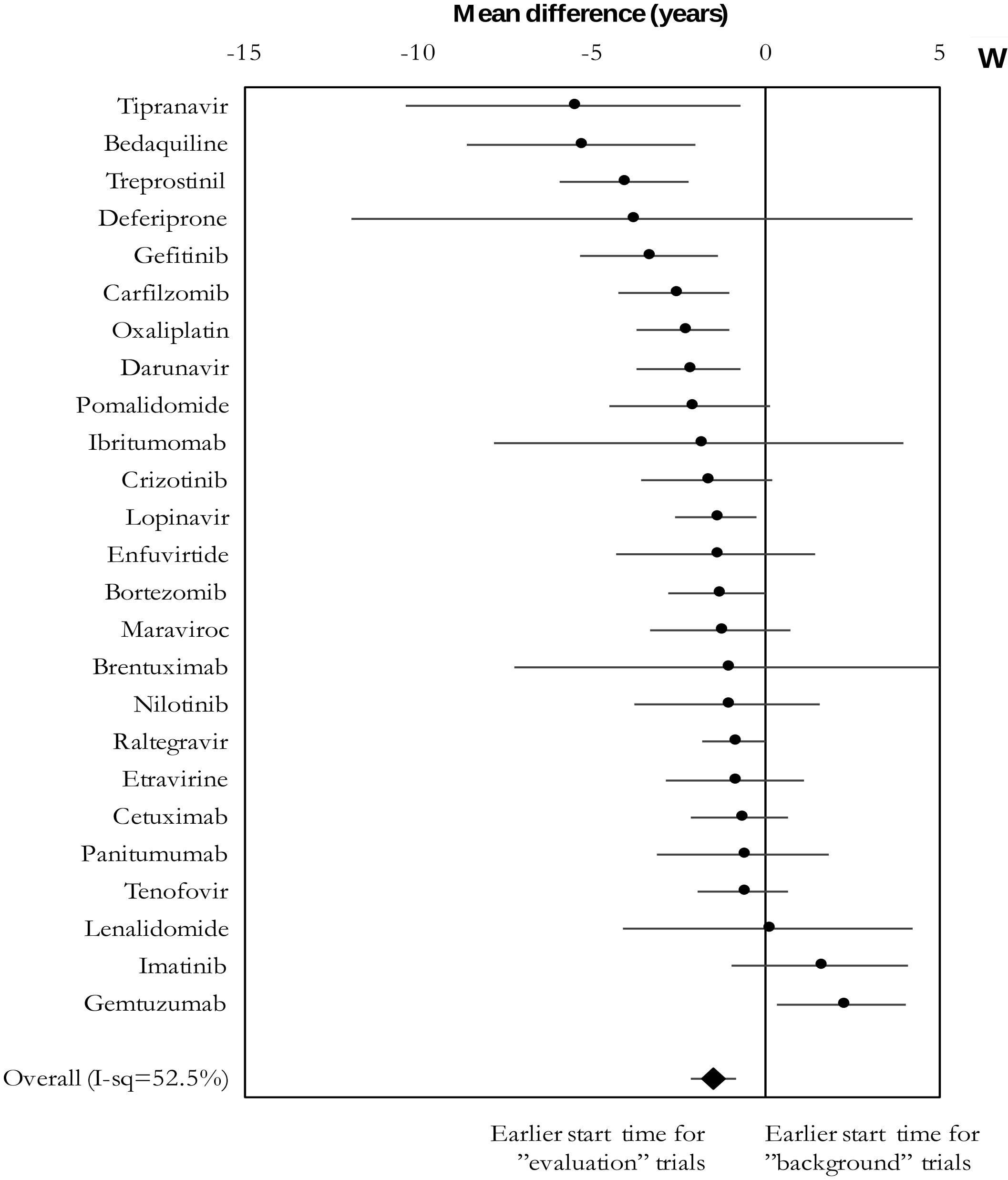
Dominant new paradigm: accelerated approvals



Accelerated approvals 2000-2013, from Naci et al. Milbank Q 2017

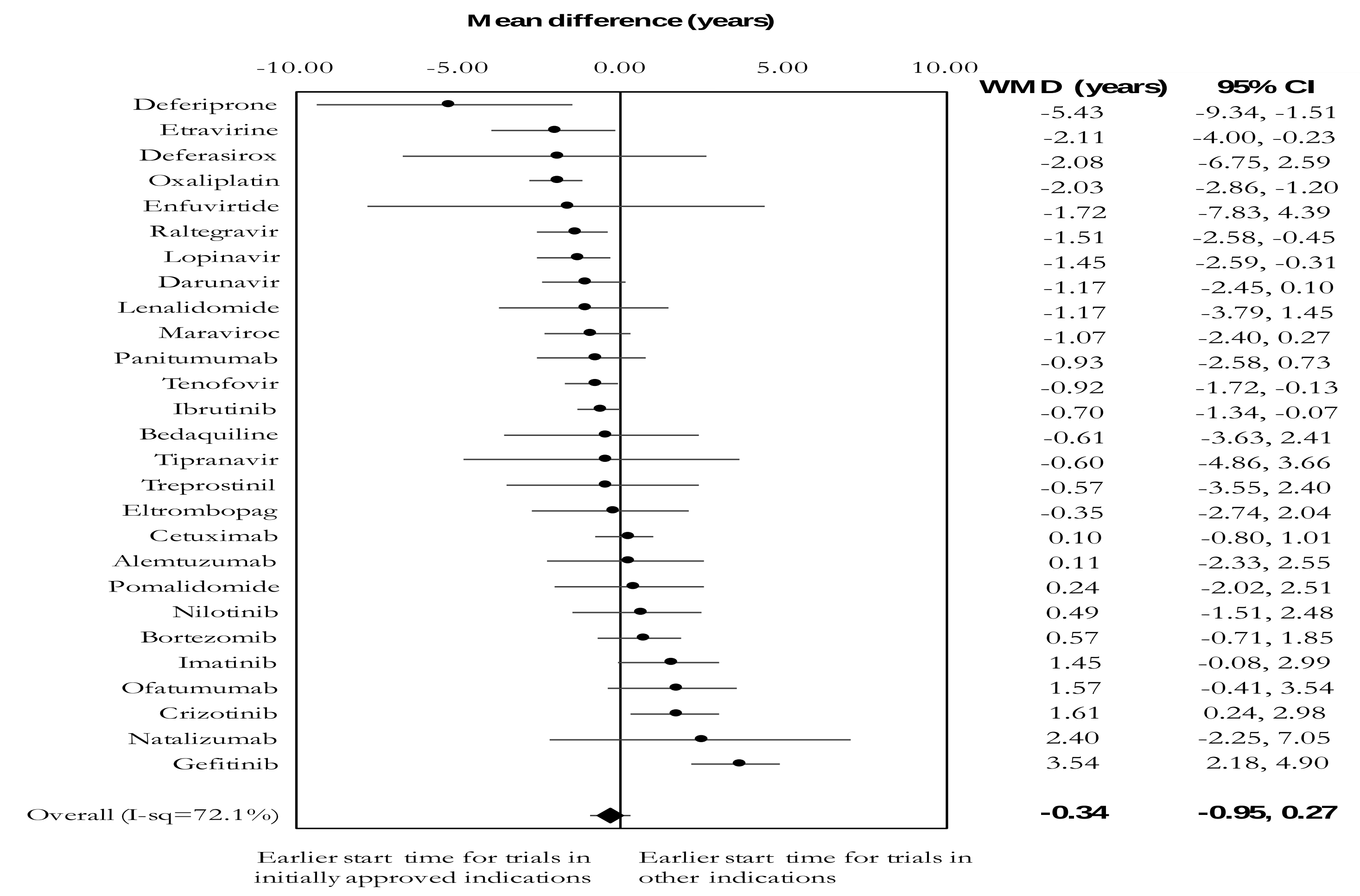
New world agenda of clinical trials: non-RCT non-indication non-evaluation





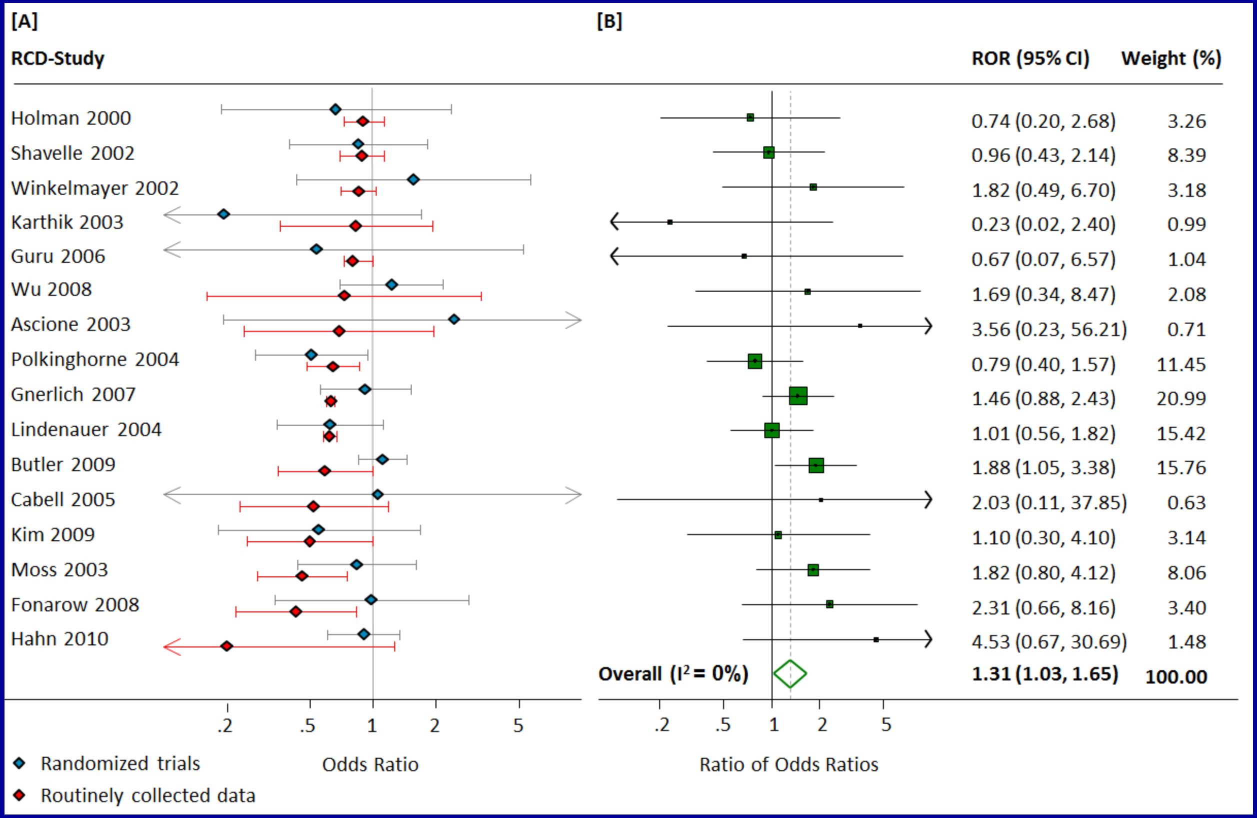
	WMD (years)	95% CI
Tipranavir	-5.55	-10.34, -0.75
Bedaquiline	-5.32	-8.59, -2.05
Treprostinil	-4.09	-5.95, -2.23
Deferiprone	-3.87	-11.94, 4.19
Gefitinib	-3.37	-5.38, -1.36
Carfilzomib	-2.65	-4.27, -1.03
Oxaliplatin	-2.38	-3.71, -1.06
Darunavir	-2.23	-3.73, -0.73
Pomalidomide	-2.19	-4.48, 0.11
Ibritumomab	-1.92	-7.80, 3.96
Crizotinib	-1.71	-3.61, 0.19
Lopinavir	-1.46	-2.62, -0.30
Enfuvirtide	-1.45	-4.31, 1.41
Bortezomib	-1.41	-2.81, 0.00
Maraviroc	-1.31	-3.31, 0.69
Brentuximab	-1.11	-7.22, 4.99
Nilotinib	-1.11	-3.78, 1.57
Raltegravir	-0.92	-1.84, 0.01
Etravirine	-0.91	-2.88, 1.07
Cetuximab	-0.75	-2.14, 0.64
Panitumumab	-0.70	-3.17, 1.78
Tenofovir	-0.65	-1.94, 0.65
Lenalidomide	0.06	-4.10, 4.22
Imatinib	1.55	-1.00, 4.09
Gemtuzumab	2.18	0.34, 4.02
Overall (I-sq=52.5%)	**-1.52**	**-2.17, -0.87**

Non-sequential steps in evidence on approved versus other indications



	WMD (years)	95% CI
Deferiprone	-5.43	-9.34, -1.51
Etravirine	-2.11	-4.00, -0.23
Deferasirox	-2.08	-6.75, 2.59
Oxaliplatin	-2.03	-2.86, -1.20
Enfuvirtide	-1.72	-7.83, 4.39
Raltegravir	-1.51	-2.58, -0.45
Lopinavir	-1.45	-2.59, -0.31
Darunavir	-1.17	-2.45, 0.10
Lenalidomide	-1.17	-3.79, 1.45
Maraviroc	-1.07	-2.40, 0.27
Panitumumab	-0.93	-2.58, 0.73
Tenofovir	-0.92	-1.72, -0.13
Ibrutinib	-0.70	-1.34, -0.07
Bedaquiline	-0.61	-3.63, 2.41
Tipranavir	-0.60	-4.86, 3.66
Treprostinil	-0.57	-3.55, 2.40
Eltrombopag	-0.35	-2.74, 2.04
Cetuximab	0.10	-0.80, 1.01
Alemtuzumab	0.11	-2.33, 2.55
Pomalidomide	0.24	-2.02, 2.51
Nilotinib	0.49	-1.51, 2.48
Bortezomib	0.57	-0.71, 1.85
Imatinib	1.45	-0.08, 2.99
Ofatumumab	1.57	-0.41, 3.54
Crizotinib	1.61	0.24, 2.98
Natalizumab	2.40	-2.25, 7.05
Gefitinib	3.54	2.18, 4.90
Overall (I-sq=72.1%)	**-0.34**	**-0.95, 0.27**

RCTs versus studies with routinely collected data



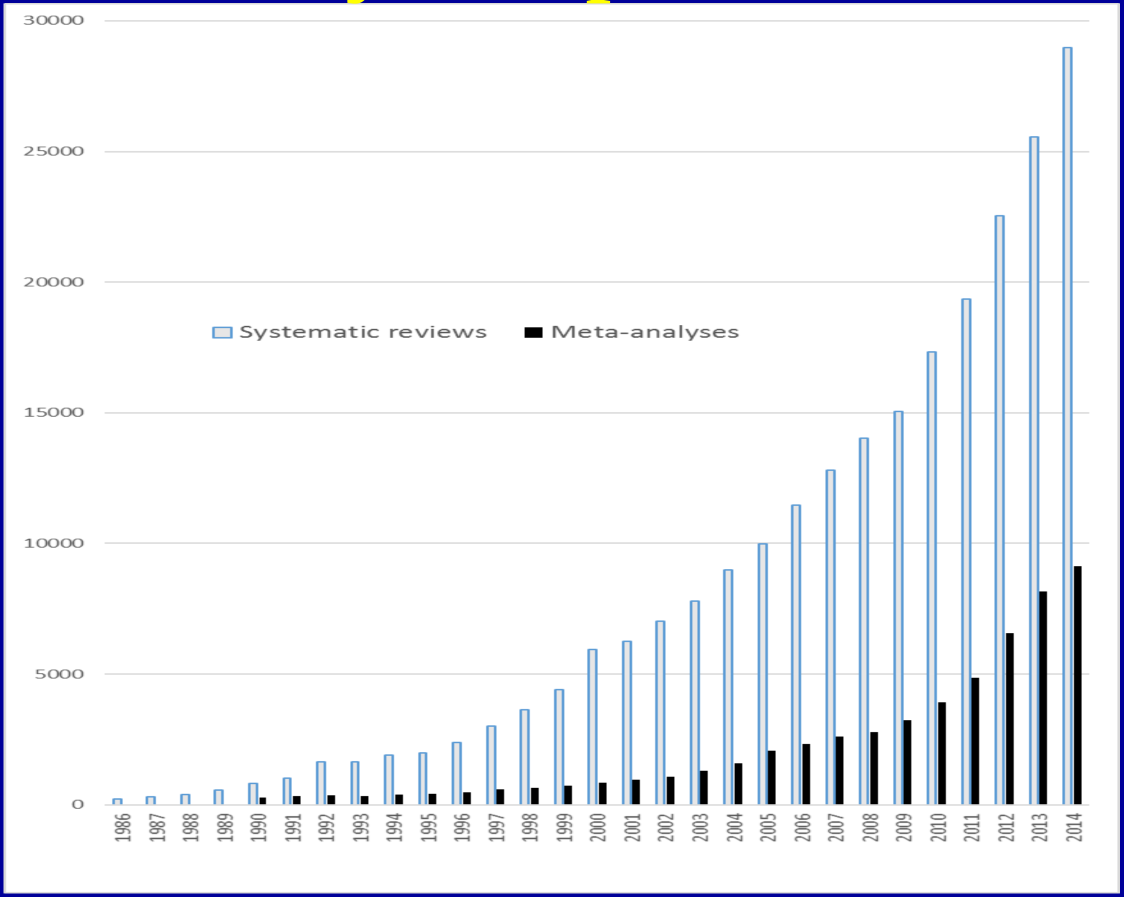
RCD-Study	ROR (95% CI)	Weight (%)
Holman 2000	0.74 (0.20, 2.68)	3.26
Shavelle 2002	0.96 (0.43, 2.14)	8.39
Winkelmayer 2002	1.82 (0.49, 6.70)	3.18
Karthik 2003	0.23 (0.02, 2.40)	0.99
Guru 2006	0.67 (0.07, 6.57)	1.04
Wu 2008	1.69 (0.34, 8.47)	2.08
Ascione 2003	3.56 (0.23, 56.21)	0.71
Polkinghorne 2004	0.79 (0.40, 1.57)	11.45
Gnerlich 2007	1.46 (0.88, 2.43)	20.99
Lindenauer 2004	1.01 (0.56, 1.82)	15.42
Butler 2009	1.88 (1.05, 3.38)	15.76
Cabell 2005	2.03 (0.11, 37.85)	0.63
Kim 2009	1.10 (0.30, 4.10)	3.14
Moss 2003	1.82 (0.80, 4.12)	8.06
Fonarow 2008	2.31 (0.66, 8.16)	3.40
Hahn 2010	4.53 (0.67, 30.69)	1.48
Overall (I² = 0%)	**1.31 (1.03, 1.65)**	**100.00**

Hemkens, Contopoulos-Ioannidis, Ioannidis, BMJ 2015

Putting the evidence together towards clinical utility: systematic reviews and meta-analyses

- As of mid-2017, there are close to 100000 published meta-analysis articles indexed in PubMed
- There are over 1000 new ones indexed every month
- There are approximately 250000 published systematic reviews in PubMed, with another 2500 new ones indexed every month

The systematic review and meta-analysis epidemic

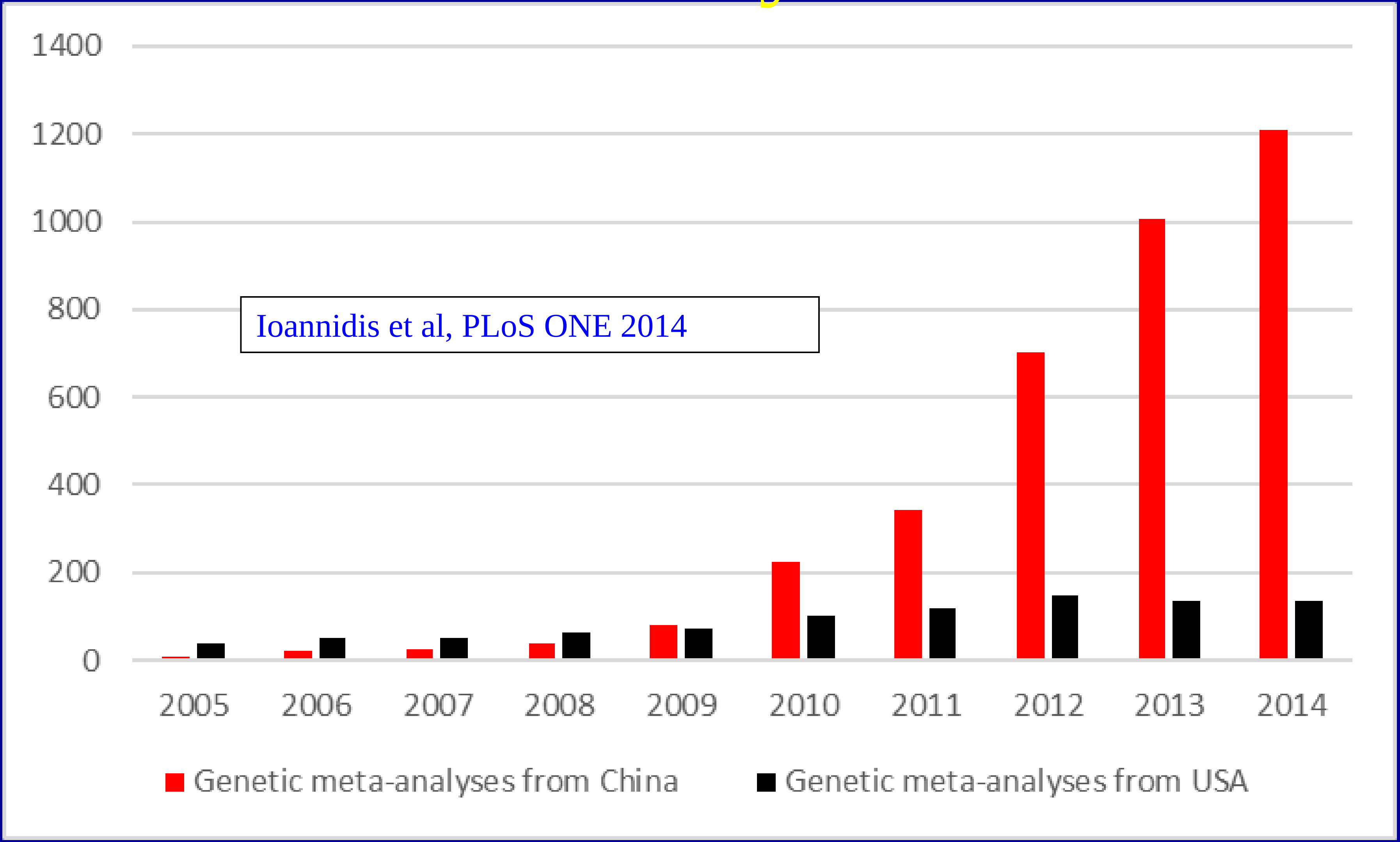


Ioannidis, Milbank Q 2016

Is this useful?

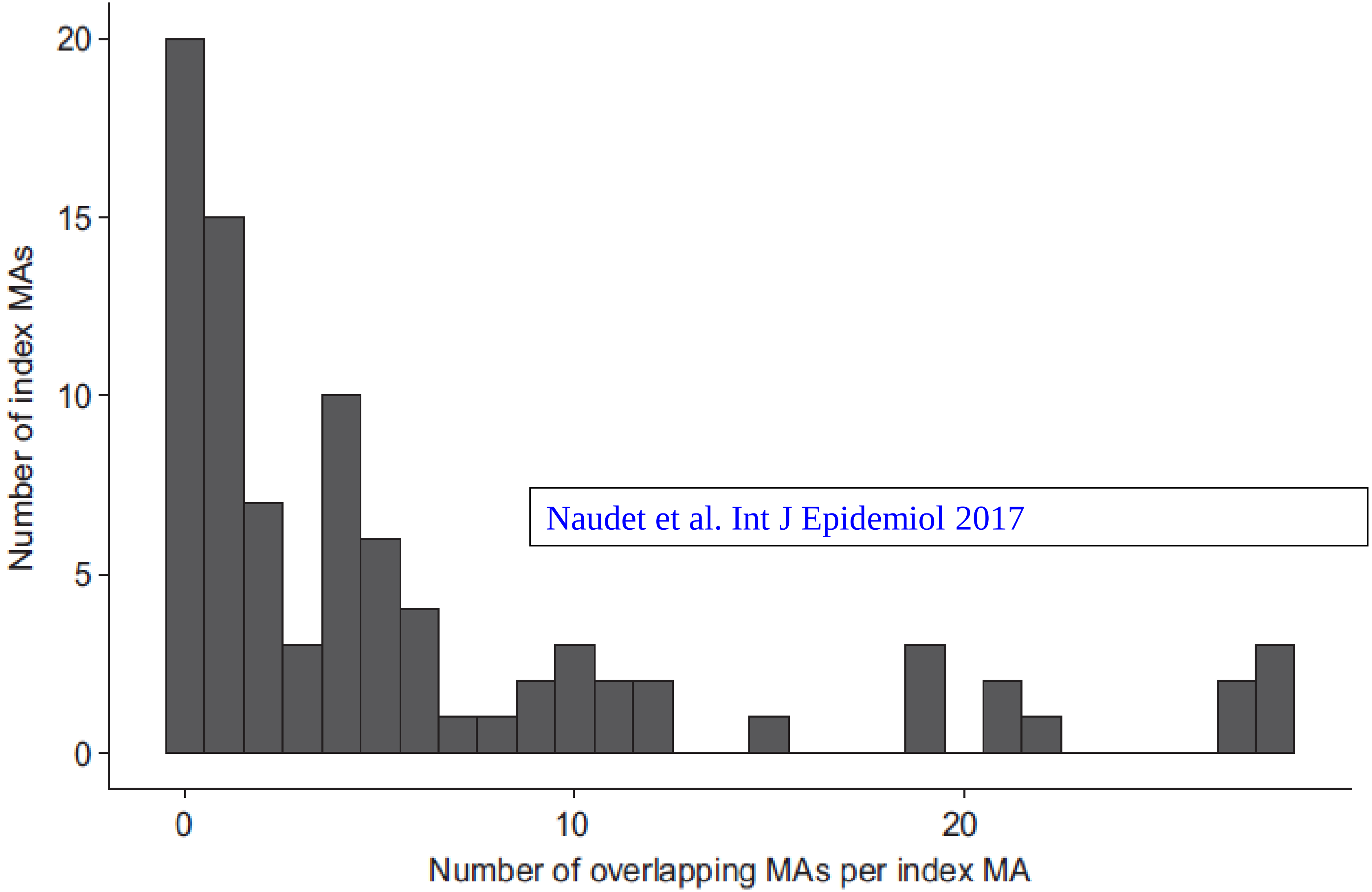
- Systematic reviews and meta-analyses have become the most powerful, influential tool of EBM
- Therefore they have been hijacked to serve various agendas
- Most systematic reviews and meta-analyses are not useful

Genetic meta-analyses from China



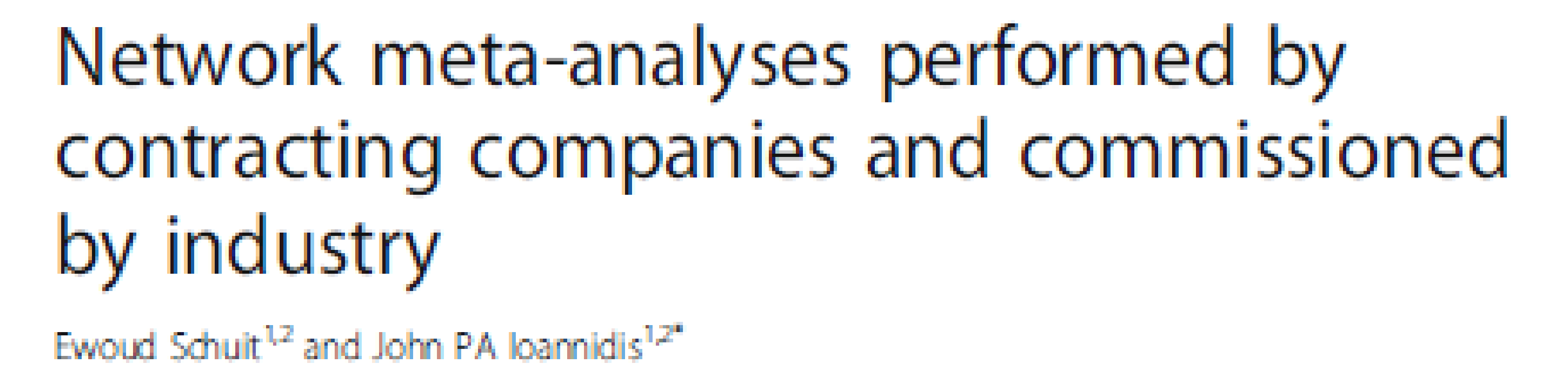
Ioannidis et al, PLoS ONE 2014

Overlapping network meta-analyses on the same topic: survey of published studies



Naudet et al. Int J Epidemiol 2017

Industry and contractors



Network meta-analyses performed by contracting companies and commissioned by industry

Ewoud Schuit[1,2] and John PA Ioannidis[1,2*]

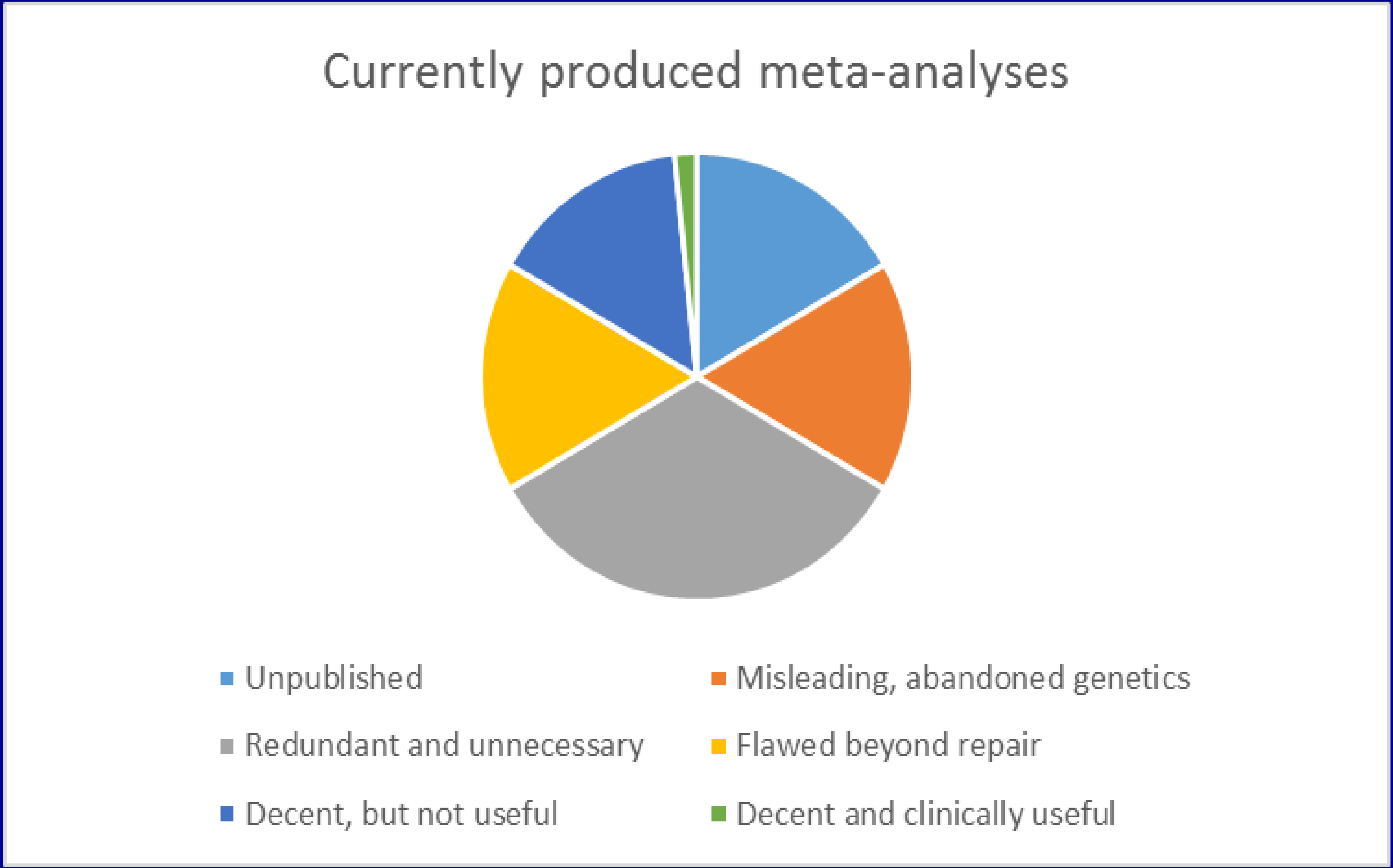
Systematic reviews as a prolific global business

- Over 100 service-offering companies perform systematic reviews
- Dozens of them perform even network meta-analyses
- Probably well over 2000 NMAs have been done by contracting for-profit companies
- Less than 20% of those have been published
- The majority of NMAs currently are done by for-profit companies hired by the industry

Schuit and Ioannidis, Syst Rev 2016

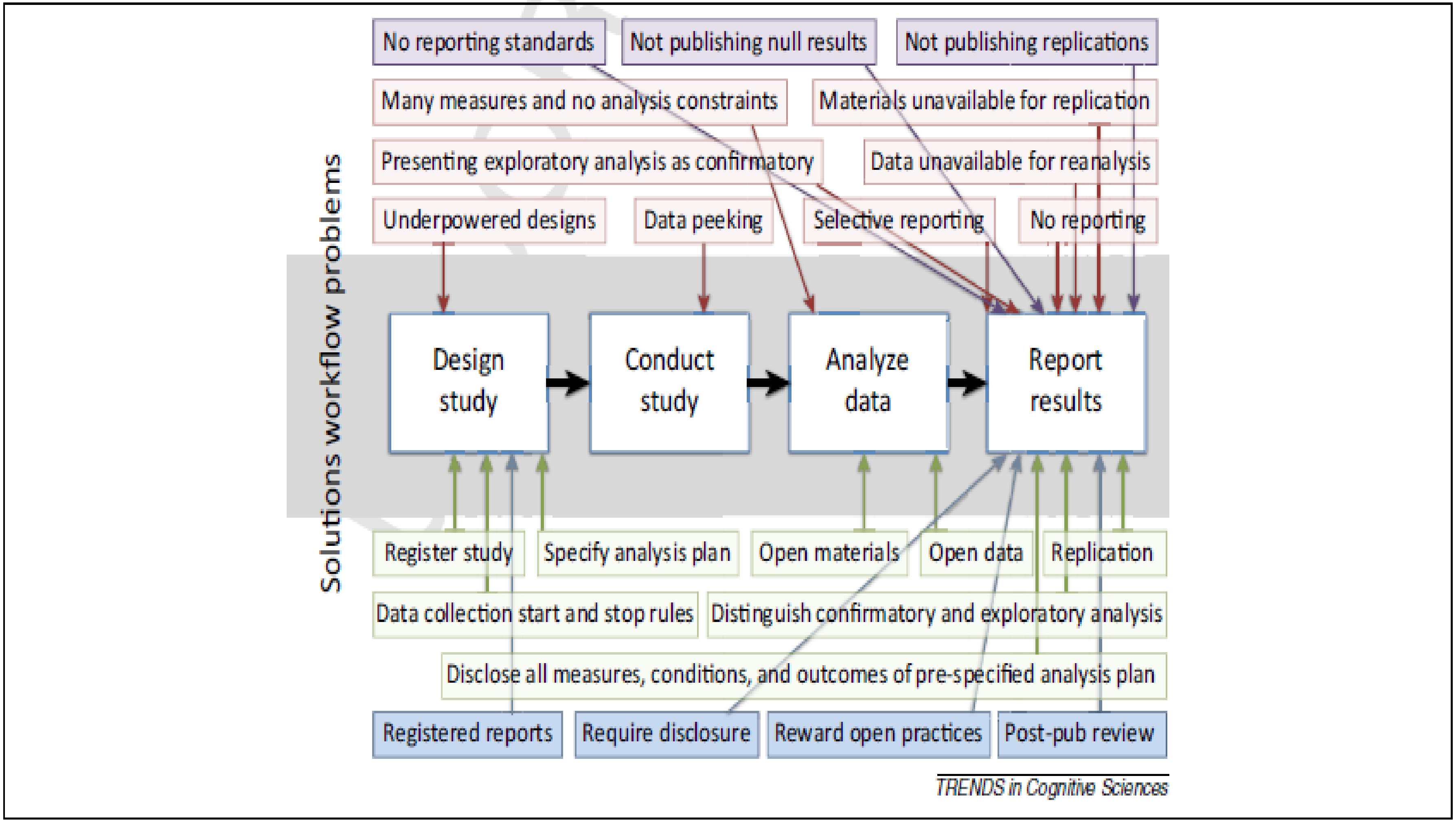
The meta-pie

(see Ioannidis, Milbank Quarterly 2016)

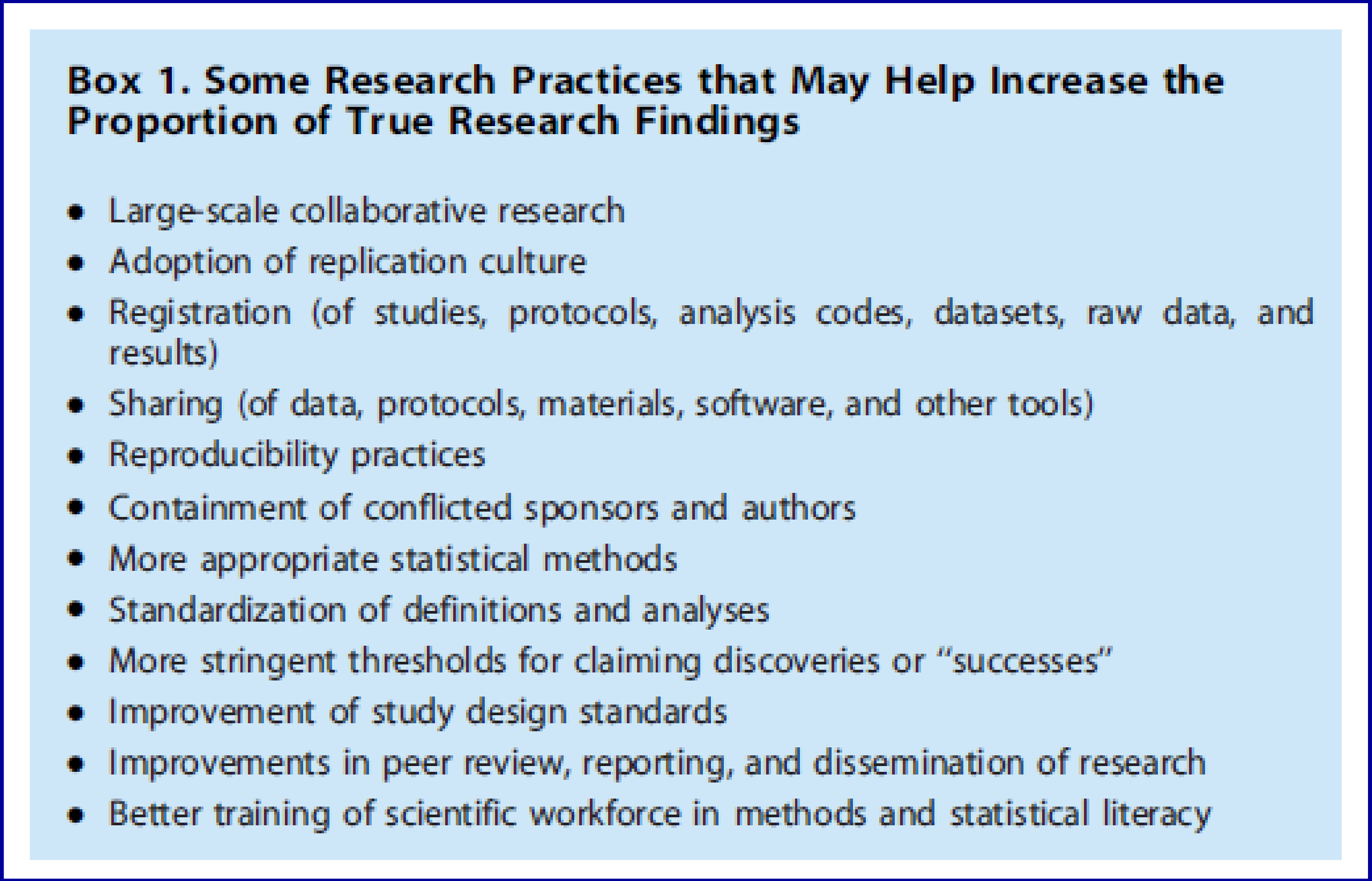


Potential solutions towards more credible and more useful research

- Some solutions have already worked in specific fields and may need to be considered in other fields as well
- Other solutions are more speculative
- Empirical evidence as to their efficacy is needed
- Seemingly effective solutions may also have collateral damages
- Do no harm


Solutions workflow problems
No reporting standards
Not publishing null results
Not publishing replications
Many measures and no analysis constraints
Materials unavailable for replication
Presenting exploratory analysis as confirmatory
Data unavailable for reanalysis
Underpowered designs
Data peeking
Selective reporting
No reporting
Design study
Conduct study
Analyze data
Report results
Register study
Specify analysis plan
Open materials
Open data
Replication
Data collection start and stop rules
Distinguish confirmatory and exploratory analysis
Disclose all measures, conditions, and outcomes of pre-specified analysis plan
Registered reports
Require disclosure
Reward open practices
Post-pub review
TRENDS in Cognitive Sciences

Ioannidis et al, Trends in Cognitive Sciences 2014

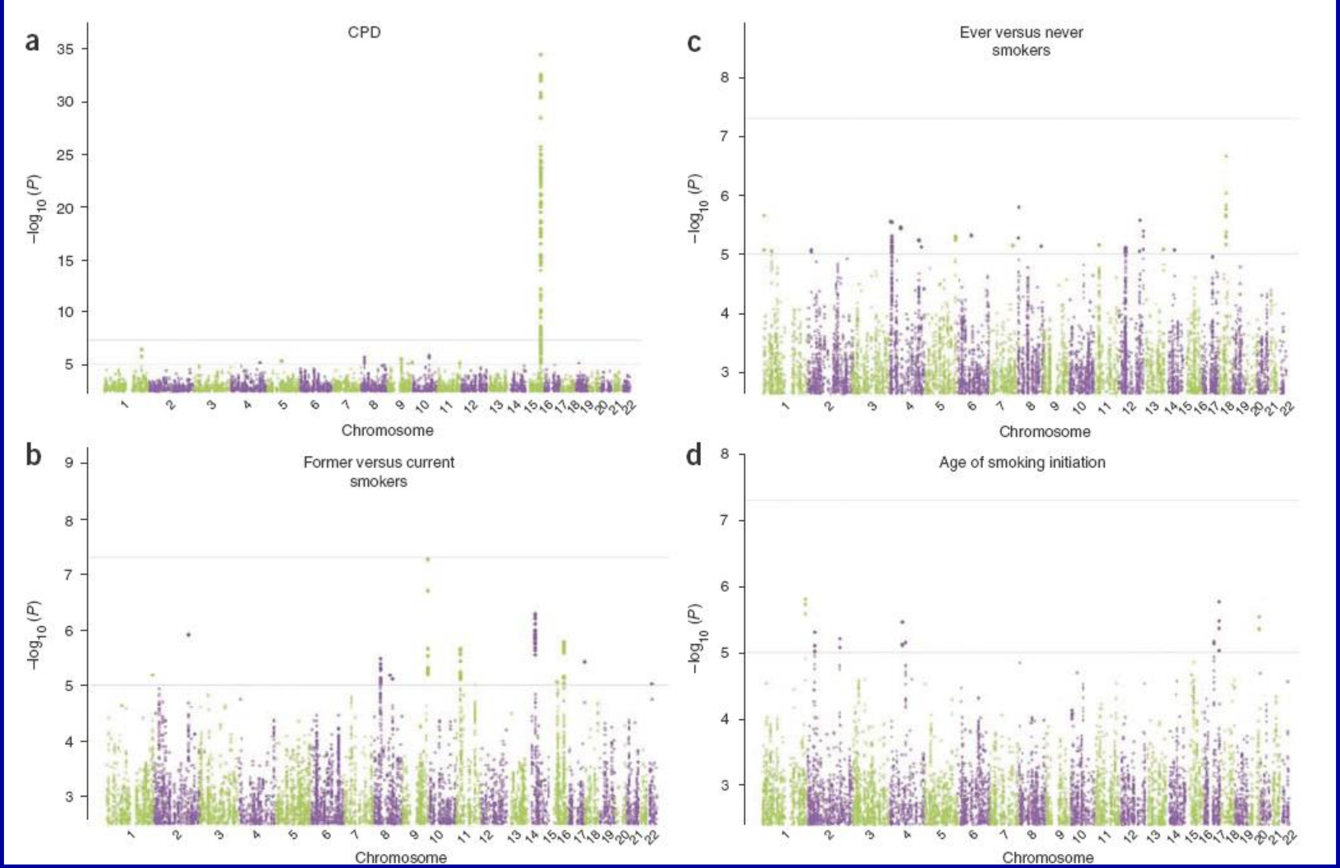


Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

- 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

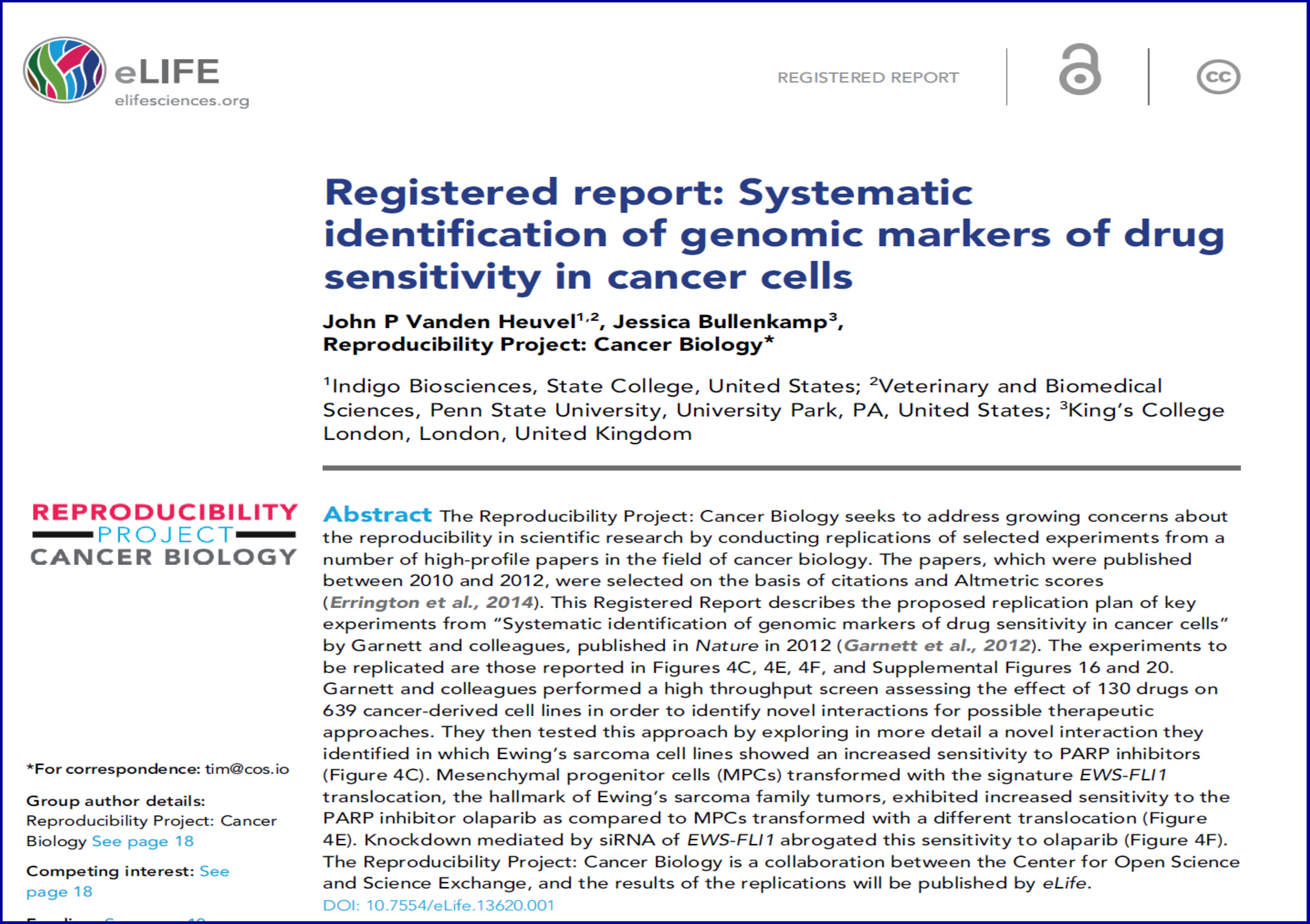
Ioannidis, PLoS Medicine 2014

Large-scale collaboration and adoption of replication culture



Levels of 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



eLIFE
elifesciences.org

REGISTERED REPORT

Registered report: Systematic identification of genomic markers of drug sensitivity in cancer cells

John P Vanden Heuvel[1,2], Jessica Bullenkamp[3], Reproducibility Project: Cancer Biology*

[1]Indigo Biosciences, State College, United States; [2]Veterinary and Biomedical Sciences, Penn State University, University Park, PA, United States; [3]King's College London, London, United Kingdom

REPRODUCIBILITY
PROJECT
CANCER BIOLOGY

Abstract The Reproducibility Project: Cancer Biology seeks to address growing concerns about the reproducibility in scientific research by conducting replications of selected experiments from a number of high-profile papers in the field of cancer biology. The papers, which were published between 2010 and 2012, were selected on the basis of citations and Altmetric scores (***Errington et al., 2014***). This Registered Report describes the proposed replication plan of key experiments from "Systematic identification of genomic markers of drug sensitivity in cancer cells" by Garnett and colleagues, published in *Nature* in 2012 (***Garnett et al., 2012***). The experiments to be replicated are those reported in Figures 4C, 4E, 4F, and Supplemental Figures 16 and 20. Garnett and colleagues performed a high throughput screen assessing the effect of 130 drugs on 639 cancer-derived cell lines in order to identify novel interactions for possible therapeutic approaches. They then tested this approach by exploring in more detail a novel interaction they identified in which Ewing's sarcoma cell lines showed an increased sensitivity to PARP inhibitors (Figure 4C). Mesenchymal progenitor cells (MPCs) transformed with the signature *EWS-FLI1* translocation, the hallmark of Ewing's sarcoma family tumors, exhibited increased sensitivity to the PARP inhibitor olaparib as compared to MPCs transformed with a different translocation (Figure 4E). Knockdown mediated by siRNA of *EWS-FLI1* abrogated this sensitivity to olaparib (Figure 4F). The Reproducibility Project: Cancer Biology is a collaboration between the Center for Open Science and Science Exchange, and the results of the replications will be published by *eLife*.

DOI: 10.7554/eLife.13620.001

***For correspondence:** tim@cos.io

Group author details: Reproducibility Project: Cancer Biology See page 18

Competing interest: See page 18

Sharing data – who, when, and how?

Doshi, Goodman, Ioannidis, TiPS 2013

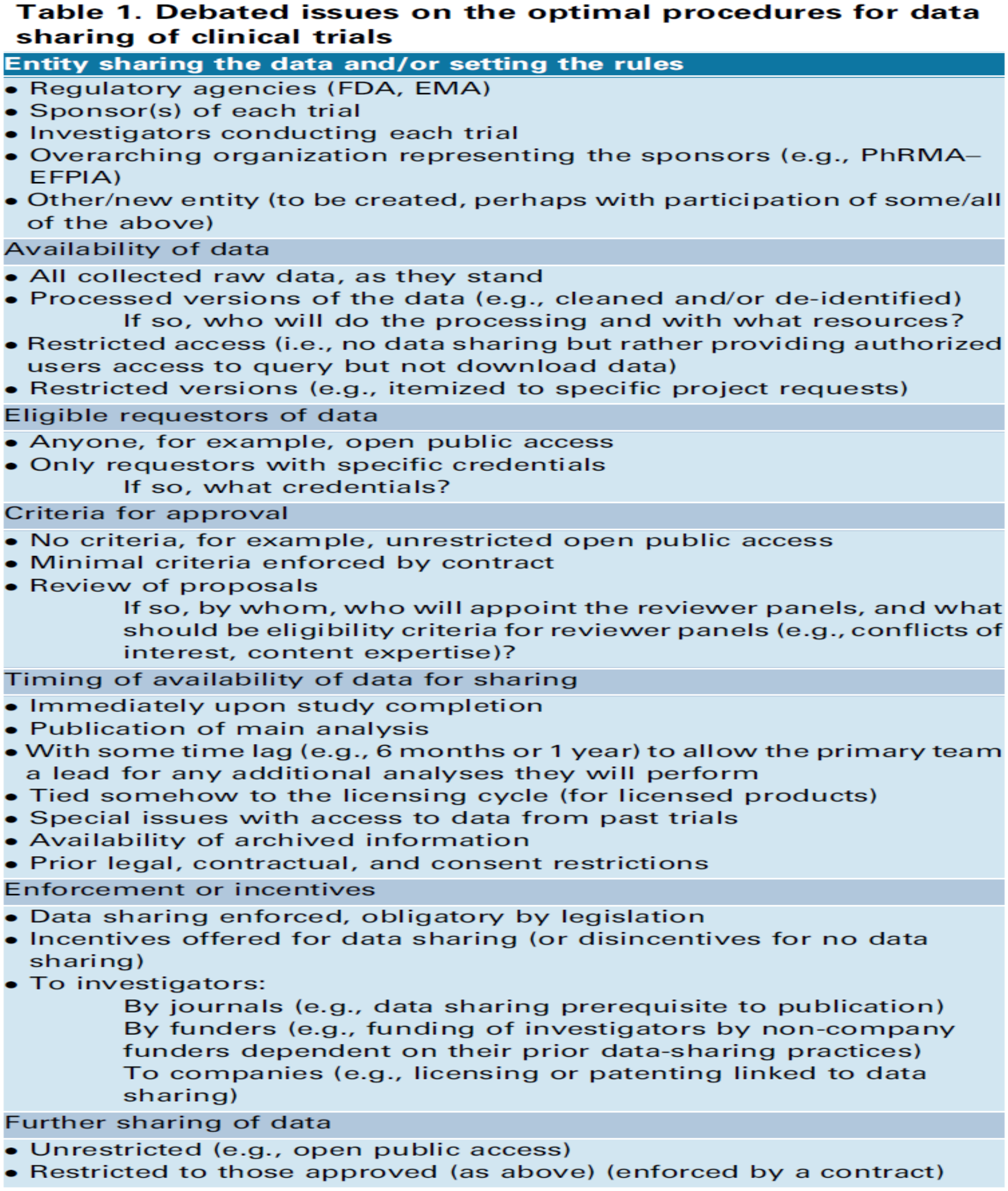


Table 1. Debated issues on the optimal procedures for data sharing of clinical trials

Entity sharing the data and/or setting the rules

- Regulatory agencies (FDA, EMA)
- Sponsor(s) of each trial
- Investigators conducting each trial
- Overarching organization representing the sponsors (e.g., PhRMA–EFPIA)
- Other/new entity (to be created, perhaps with participation of some/all of the above)

Availability of data

- All collected raw data, as they stand
- Processed versions of the data (e.g., cleaned and/or de-identified)
  - If so, who will do the processing and with what resources?
- Restricted access (i.e., no data sharing but rather providing authorized users access to query but not download data)
- Restricted versions (e.g., itemized to specific project requests)

Eligible requestors of data

- Anyone, for example, open public access
- Only requestors with specific credentials
  - If so, what credentials?

Criteria for approval

- No criteria, for example, unrestricted open public access
- Minimal criteria enforced by contract
- Review of proposals
  - If so, by whom, who will appoint the reviewer panels, and what should be eligibility criteria for reviewer panels (e.g., conflicts of interest, content expertise)?

Timing of availability of data for sharing

- Immediately upon study completion
- Publication of main analysis
- With some time lag (e.g., 6 months or 1 year) to allow the primary team a lead for any additional analyses they will perform
- Tied somehow to the licensing cycle (for licensed products)
- Special issues with access to data from past trials
- Availability of archived information
- Prior legal, contractual, and consent restrictions

Enforcement or incentives

- Data sharing enforced, obligatory by legislation
- Incentives offered for data sharing (or disincentives for no data sharing)
- To investigators:
  - By journals (e.g., data sharing prerequisite to publication)
  - By funders (e.g., funding of investigators by non-company funders dependent on their prior data-sharing practices)
  - To companies (e.g., licensing or patenting linked to data sharing)

Further sharing of data

- Unrestricted (e.g., open public access)
- Restricted to those approved (as above) (enforced by a contract)

PLOS | BIOLOGY

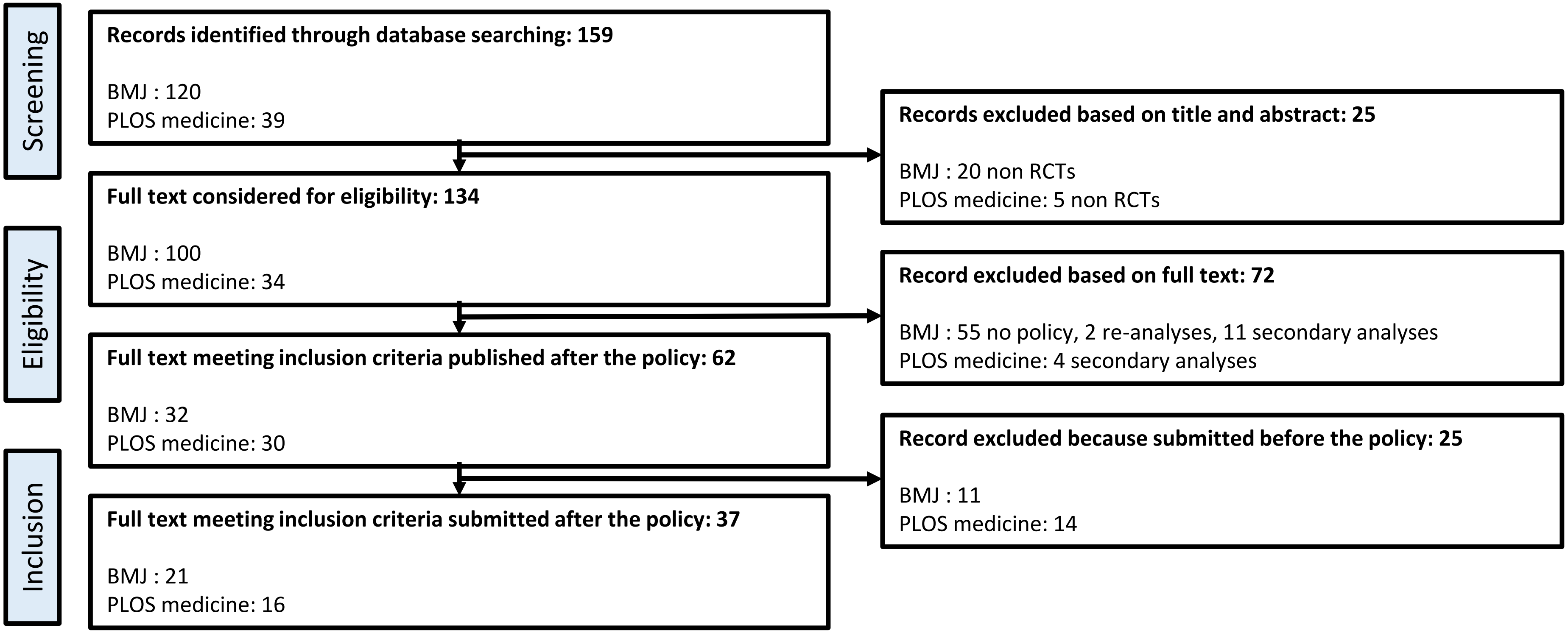
META-RESEARCH ARTICLE

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.

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



Screening

Records identified through database searching: 159

BMJ : 120
PLOS medicine: 39

Records excluded based on title and abstract: 25

BMJ : 20 non RCTs
PLOS medicine: 5 non RCTs

Eligibility

Full text considered for eligibility: 134

BMJ : 100
PLOS medicine: 34

Record excluded based on full text: 72

BMJ : 55 no policy, 2 re-analyses, 11 secondary analyses
PLOS medicine: 4 secondary analyses

Full text meeting inclusion criteria published after the policy: 62

BMJ : 32
PLOS medicine: 30

Record excluded because submitted before the policy: 25

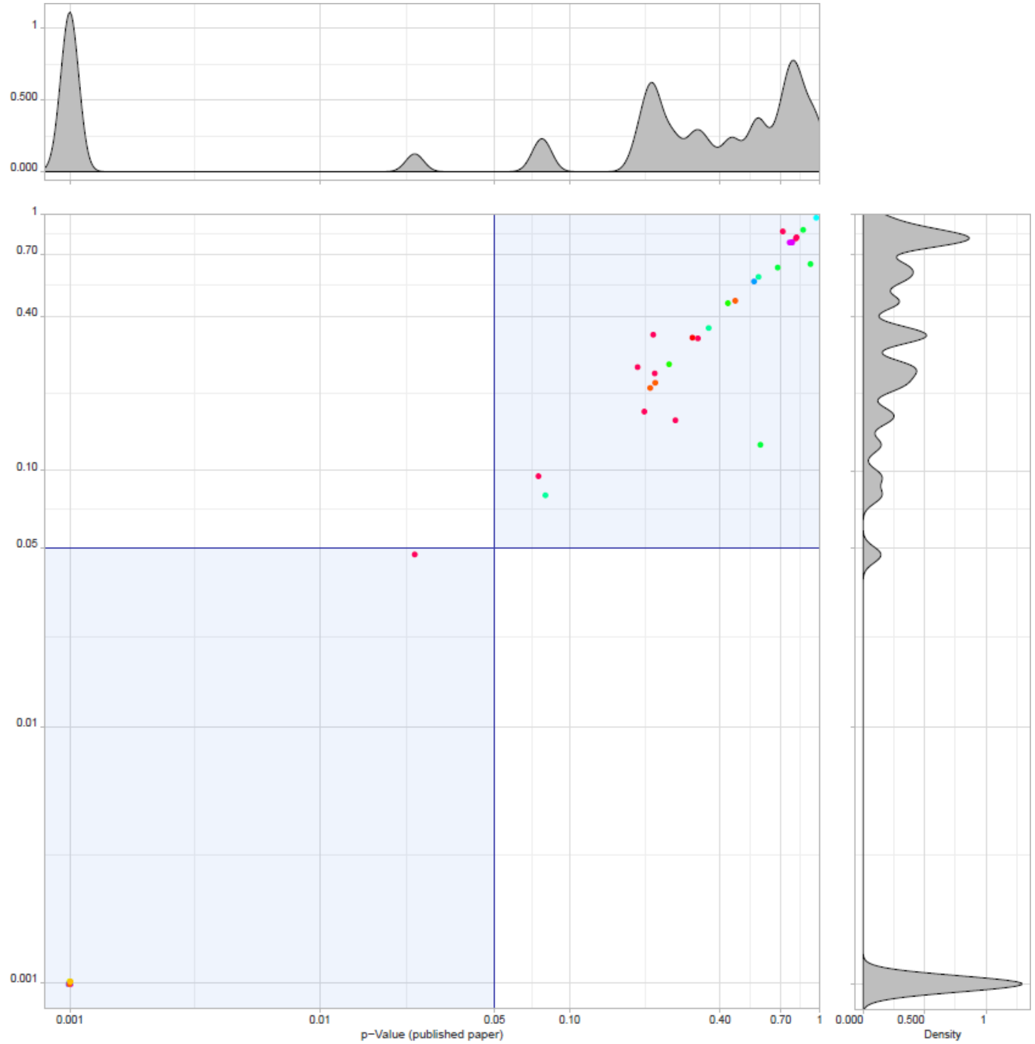
BMJ : 11
PLOS medicine: 14

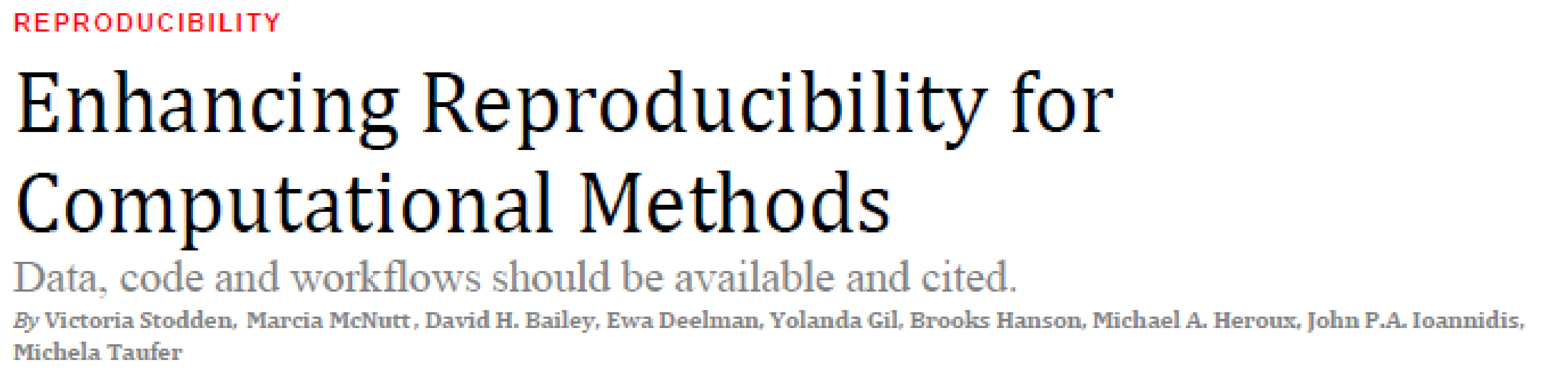
Inclusion

Full text meeting inclusion criteria submitted after the policy: 37

BMJ : 21
PLOS medicine: 16

Naudet et al, submitted


1
0.500
0.000
1
0.70
0.40
0.10
0.05
0.01
0.001
0.001
0.01
0.05
0.10
0.40
0.70
1
p−Value (published paper)
0.000
0.500
1
Density



REPRODUCIBILITY

Enhancing Reproducibility for Computational Methods

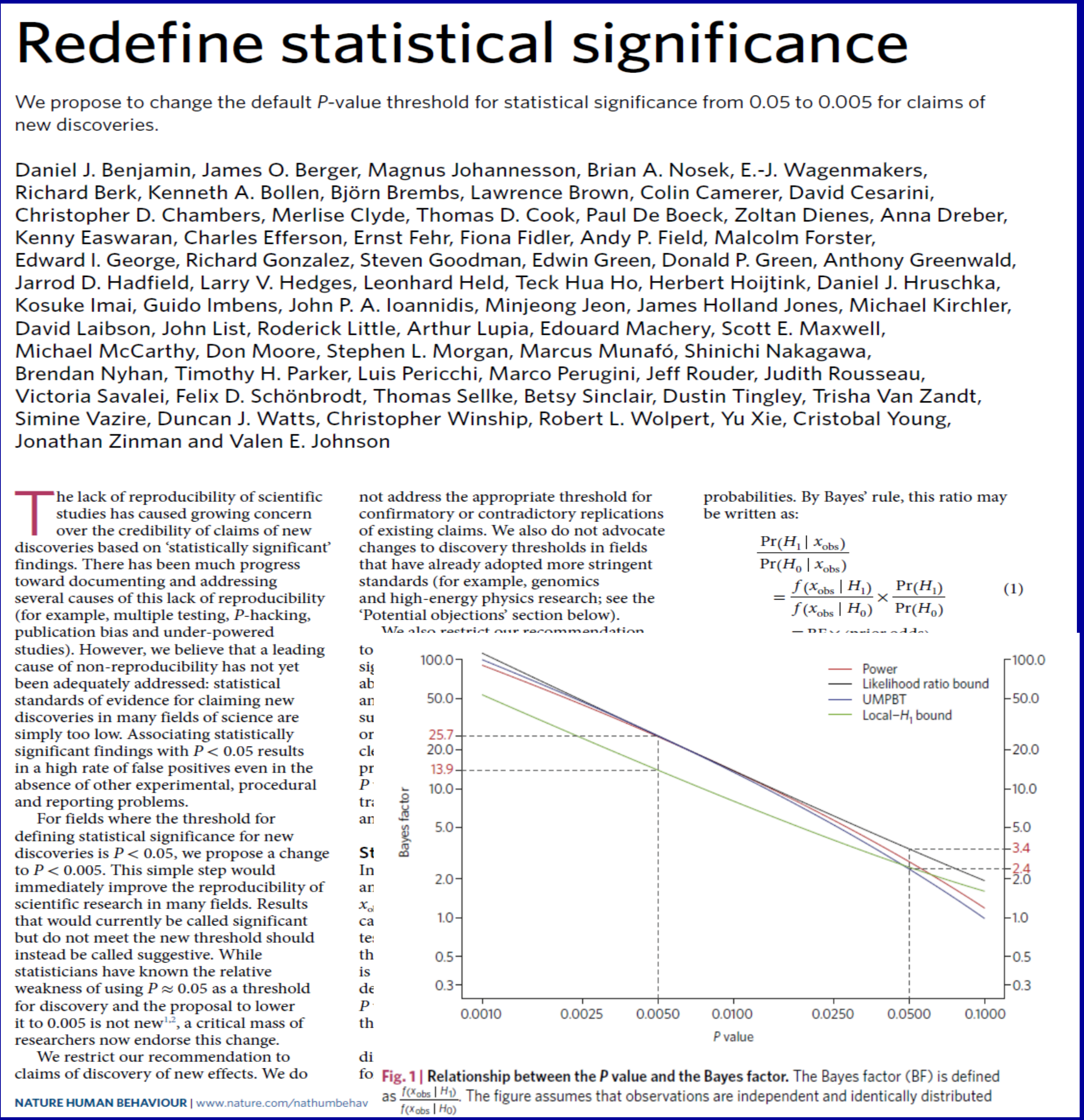
Data, code and workflows should be available and cited.

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

Better statistics and methods

- Transparent (registered?) statistical analysis plans
- Statistical training and improved literacy/numeracy of scientific workforce
- Better study designs
- Standard features: e.g. randomization and blinding of investigators in animal experiments
- What role for design/conduct checklists?



Redefine statistical significance

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

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

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

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

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

We also restrict our recommendation

to
sig
ab
an
su
or
cl
pr
P
tra
an

St
In
an
x_o
ca
te
th
is
de
P
th

di
fo

probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{\text{obs}})}{\Pr(H_0 \mid x_{\text{obs}})} = \frac{f(x_{\text{obs}} \mid H_1)}{f(x_{\text{obs}} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \quad (1)$$

$$= \text{BF} \times (\text{prior odds})$$

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

Nature Human Behavior September 1, 2017

frontiers
in Human Neuroscience

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

Check for updates

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

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

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

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

OPEN ACCESS

Edited by:

Is NHST a good choice for:

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

Is it up to institutional changes?

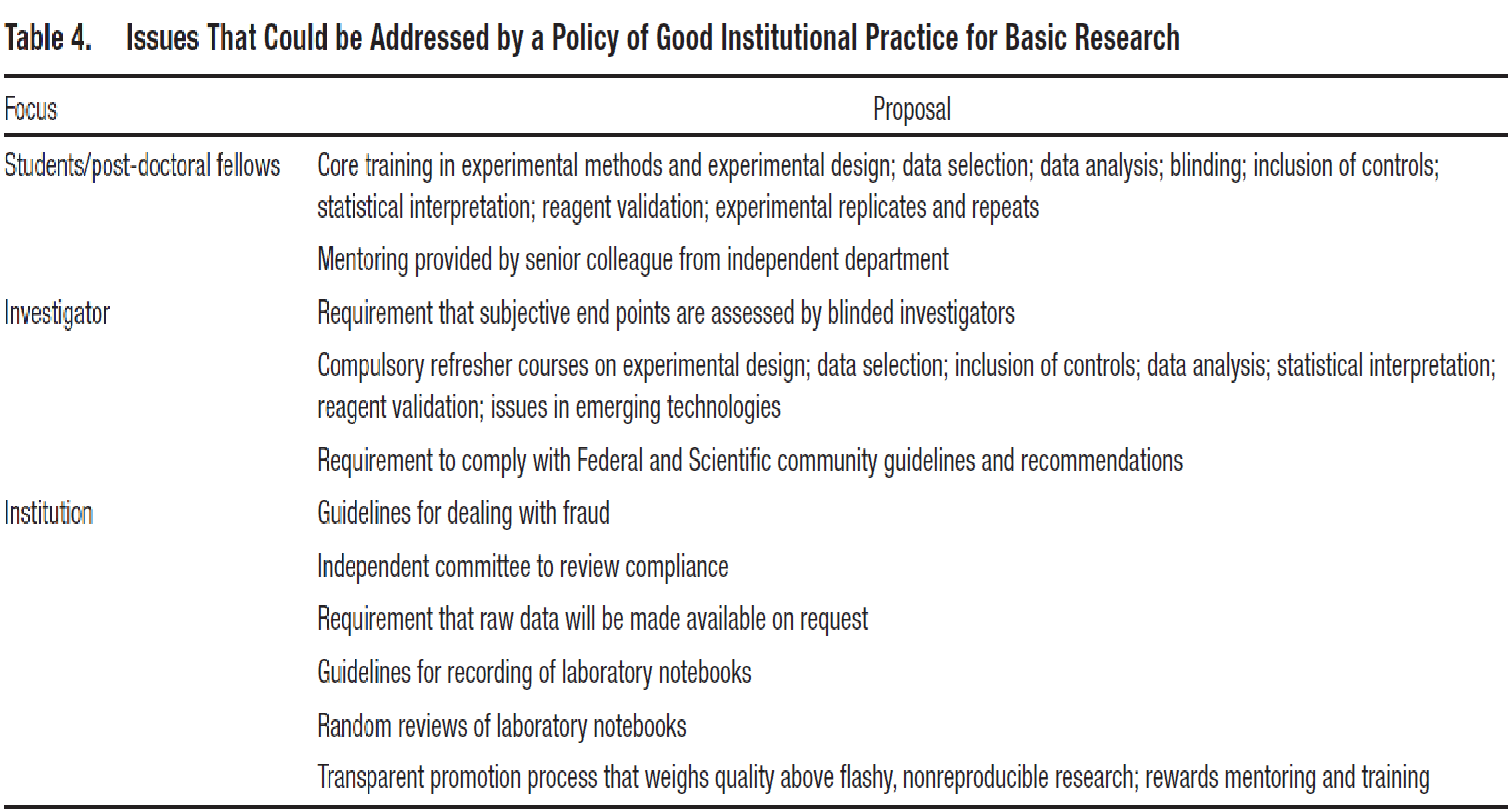
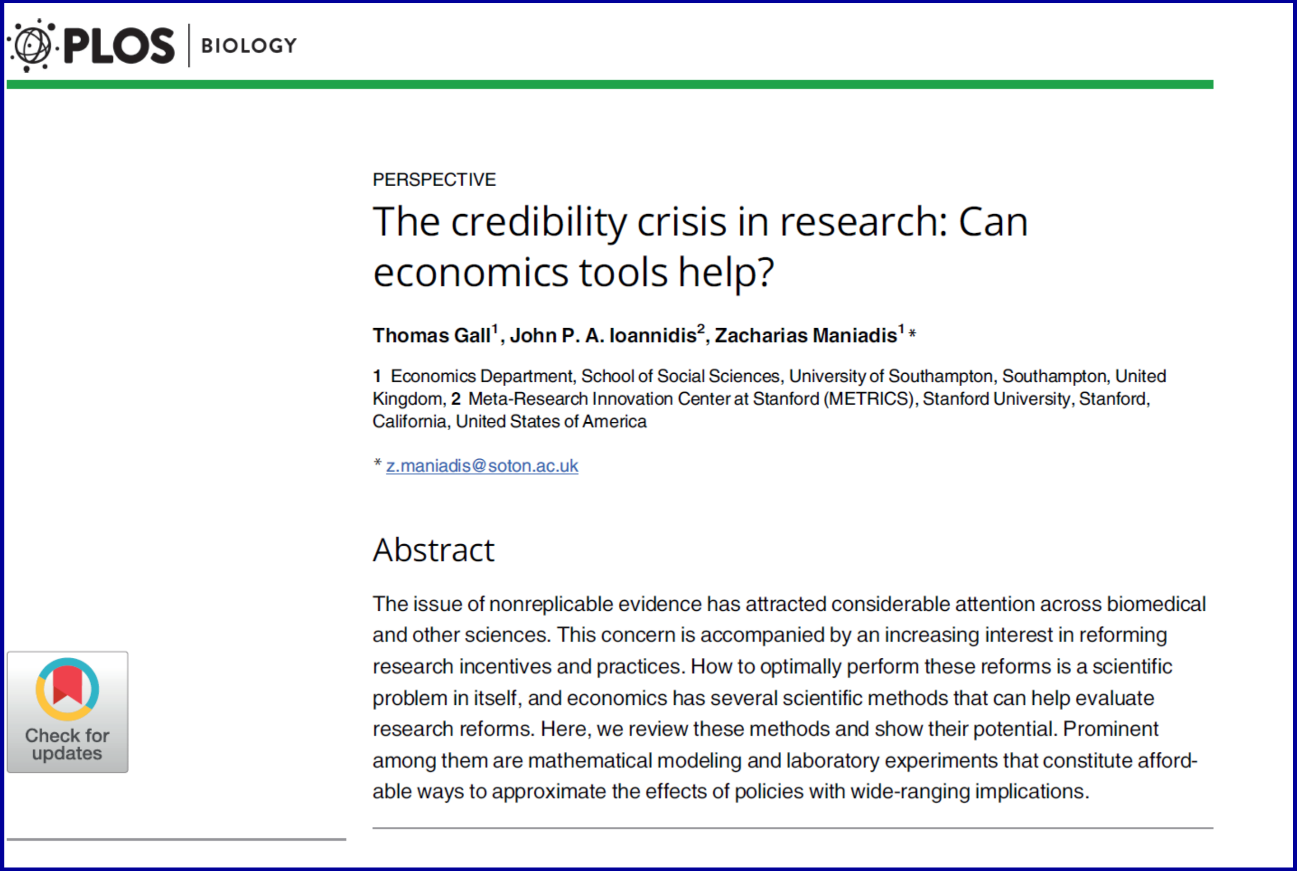


Table 4. Issues That Could be Addressed by a Policy of Good Institutional Practice for Basic Research

Focus	Proposal
Students/post-doctoral fellows	Core training in experimental methods and experimental design; data selection; data analysis; blinding; inclusion of controls; statistical interpretation; reagent validation; experimental replicates and repeats
	Mentoring provided by senior colleague from independent department
Investigator	Requirement that subjective end points are assessed by blinded investigators
	Compulsory refresher courses on experimental design; data selection; inclusion of controls; data analysis; statistical interpretation; reagent validation; issues in emerging technologies
	Requirement to comply with Federal and Scientific community guidelines and recommendations
Institution	Guidelines for dealing with fraud
	Independent committee to review compliance
	Requirement that raw data will be made available on request
	Guidelines for recording of laboratory notebooks
	Random reviews of laboratory notebooks
	Transparent promotion process that weighs quality above flashy, nonreproducible research; rewards mentoring and training

Begley and Ioannidis, Circulation Research 2015

Modeling and modeling plus experimentation



PLOS | BIOLOGY

PERSPECTIVE

The credibility crisis in research: Can economics tools help?

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

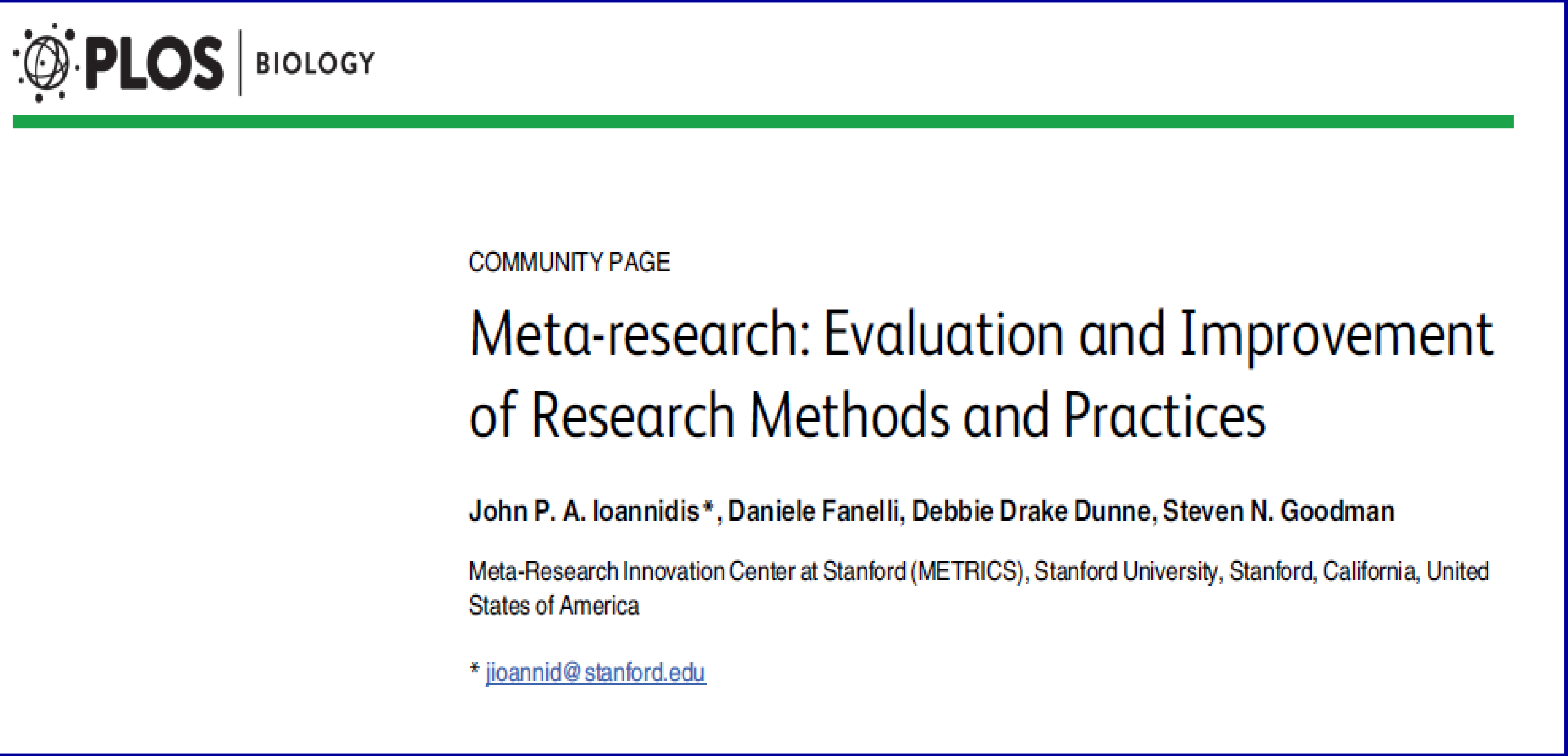
1 Economics Department, School of Social Sciences, University of Southampton, Southampton, United Kingdom, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States of America

* z.maniadis@soton.ac.uk

Check for updates

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.



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

Re-engineering the reward system

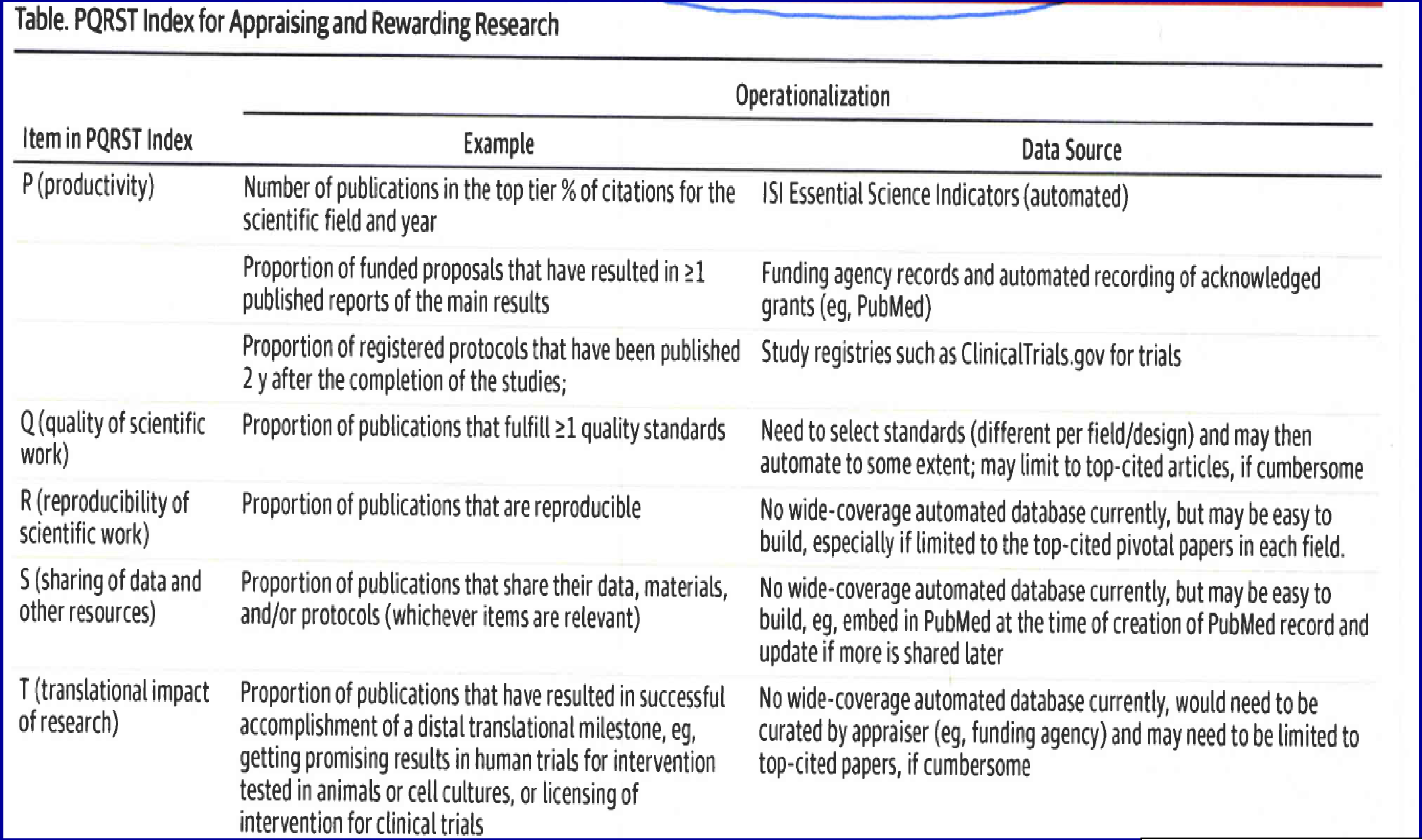
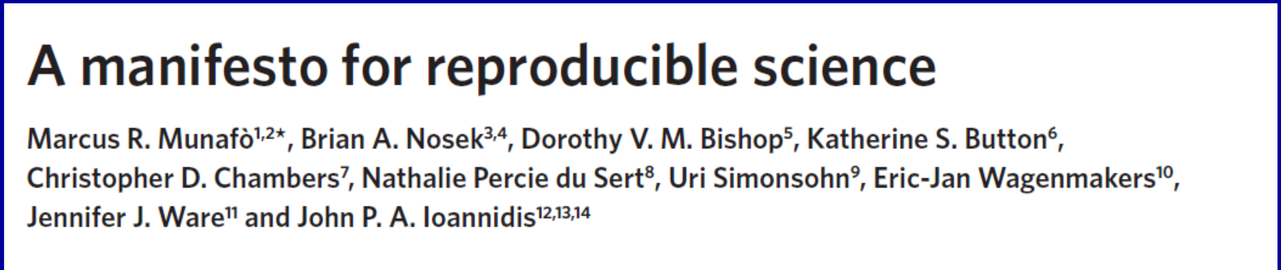


Table. PQRST Index for Appraising and Rewarding Research

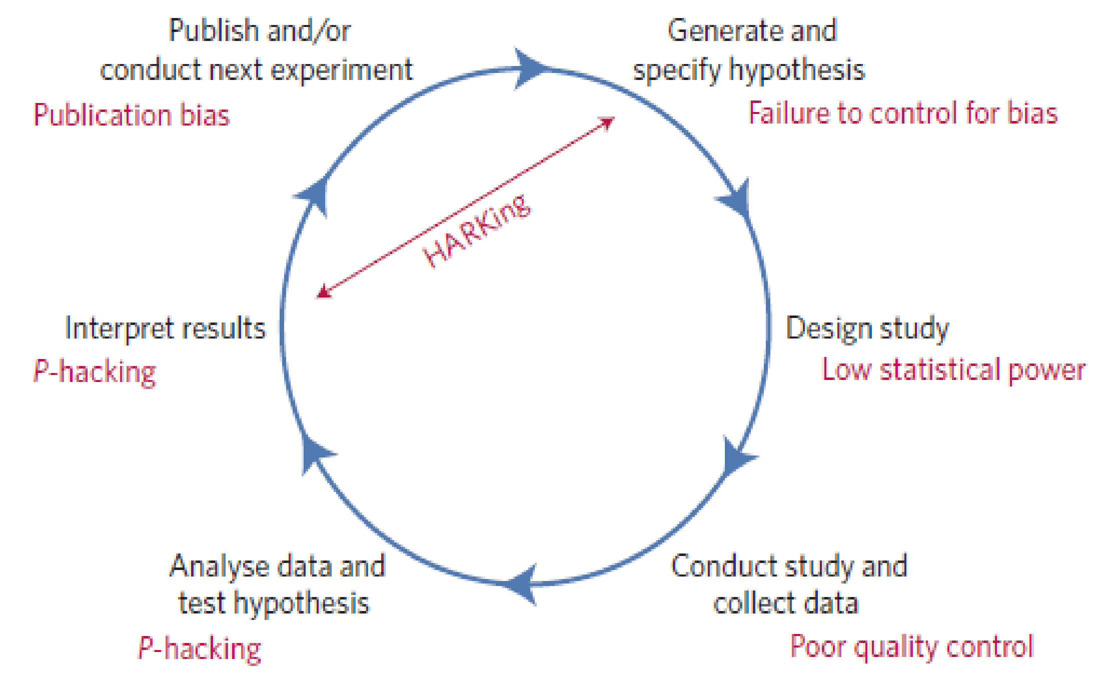
Item in PQRST Index	Operationalization	
	Example	Data Source
P (productivity)	Number of publications in the top tier % of citations for the scientific field and year	ISI Essential Science Indicators (automated)
	Proportion of funded proposals that have resulted in ≥1 published reports of the main results	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 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]



Publish and/or conduct next experiment

Publication bias

Generate and specify hypothesis

Failure to control for bias

HARKing

Interpret results

P-hacking

Design study

Low statistical power

Analyse data and test hypothesis

P-hacking

Conduct study and collect data

Poor quality control

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

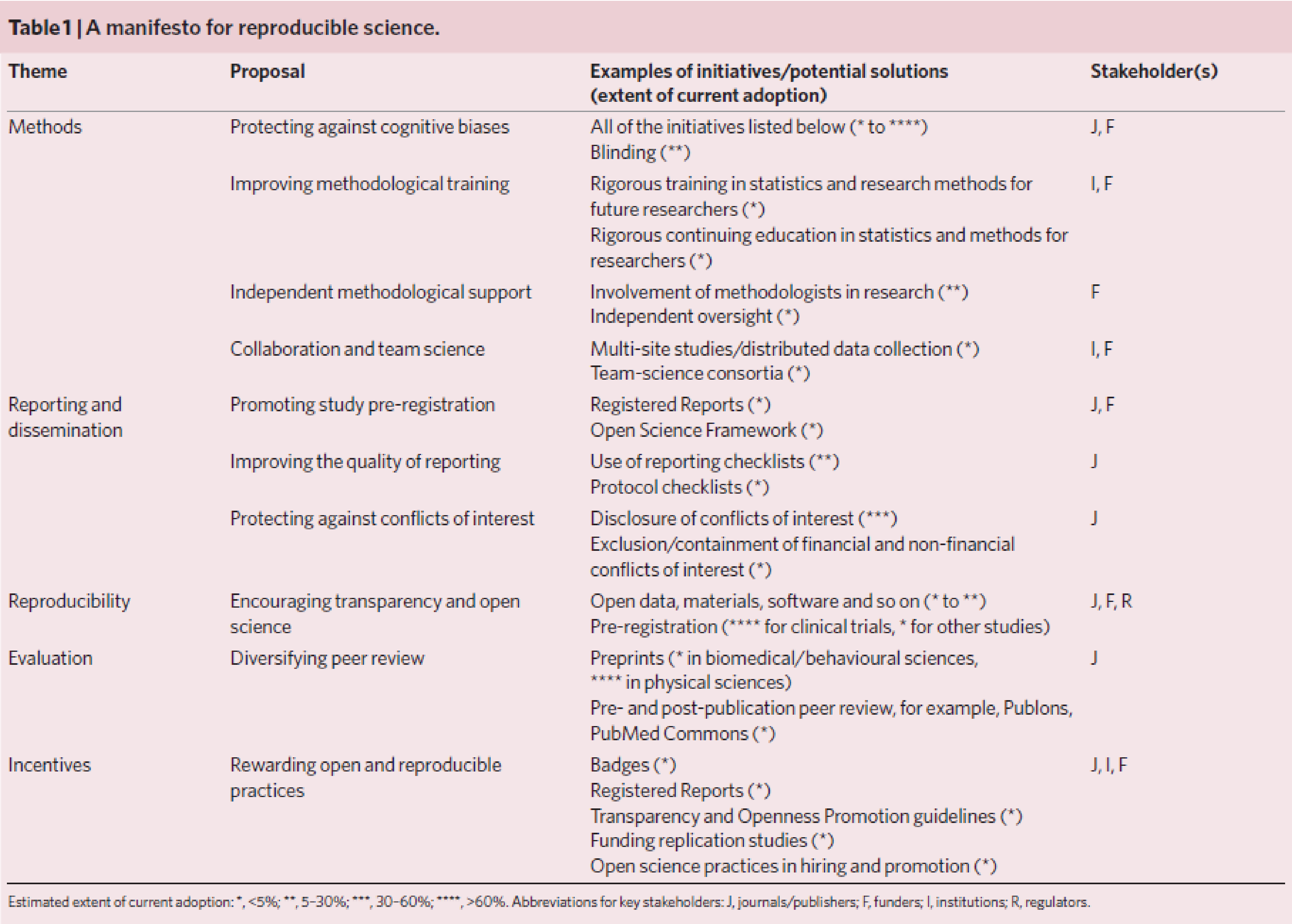


Table 1 | A manifesto for reproducible science.

Theme	Proposal	Examples of initiatives/potential solutions (extent of current adoption)	Stakeholder(s)
Methods	Protecting against cognitive biases	All of the initiatives listed below (* to ****) Blinding (**)	J, F
	Improving methodological training	Rigorous training in statistics and research methods for future researchers (*) Rigorous continuing education in statistics and methods for researchers (*)	I, F
	Independent methodological support	Involvement of methodologists in research (**) Independent oversight (*)	F
	Collaboration and team science	Multi-site studies/distributed data collection (*) Team-science consortia (*)	I, F
Reporting and dissemination	Promoting study pre-registration	Registered Reports (*) Open Science Framework (*)	J, F
	Improving the quality of reporting	Use of reporting checklists (**) Protocol checklists (*)	J
	Protecting against conflicts of interest	Disclosure of conflicts of interest (***) Exclusion/containment of financial and non-financial conflicts of interest (*)	J
Reproducibility	Encouraging transparency and open science	Open data, materials, software and so on (* to **) Pre-registration (**** for clinical trials, * for other studies)	J, F, R
Evaluation	Diversifying peer review	Preprints (* in biomedical/behavioural sciences, **** in physical sciences) Pre- and post-publication peer review, for example, Publons, PubMed Commons (*)	J
Incentives	Rewarding open and reproducible practices	Badges (*) Registered Reports (*) Transparency and Openness Promotion guidelines (*) Funding replication studies (*) Open science practices in hiring and promotion (*)	J, I, F

Estimated extent of current adoption: *, <5%; **, 5–30%; ***, 30–60%; ****, >60%. Abbreviations for key stakeholders: J, journals/publishers; F, funders; I, institutions; R, regulators.

Understand and align interests of stakeholders

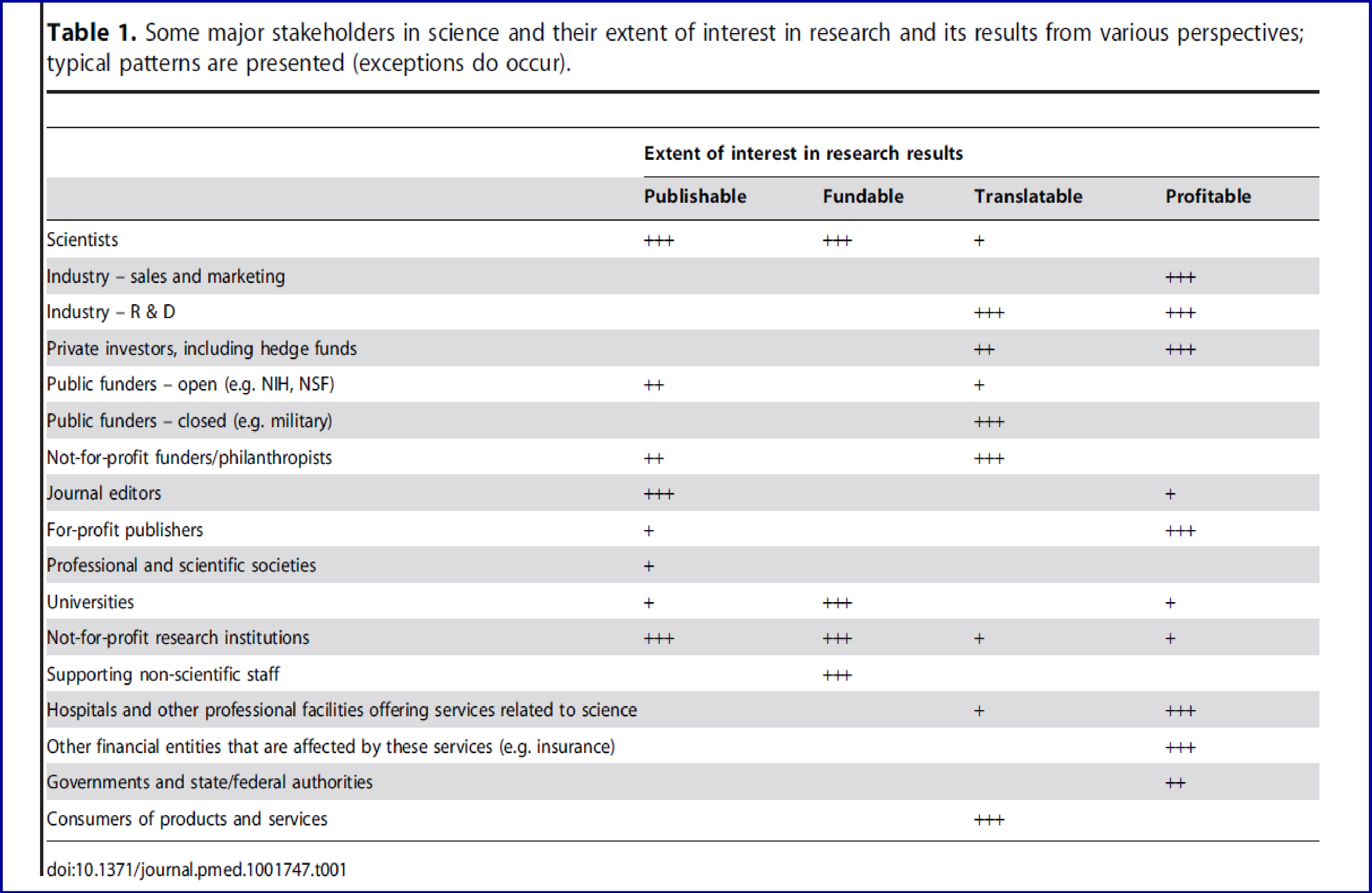


Table 1. Some major stakeholders in science and their extent of interest in research and its results from various perspectives; typical patterns are presented (exceptions do occur).

	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

…Στο ανακαινισμένο θέατρο θα στηθούν το απόγευμα τα επτά μικρόφωνα για τους απόντες ομιλητές. Μετά τους μονομάχους, θα έρθουν οι οργανοπαίχτες κι έπειτα οι σύνεδροι επιστήμονες παραπαίοντας στον περίπατο των κυπαρισσιών. Μόνο η σαύρα ξέρει τελικά να ορθώνει κεφάλι, κι όχι, φυσικά, δεν είναι ο άνθρωπος που θα ρυμουλκήσει τη φύση που εγκλωβίστηκε στους νόμους της.

…The renovated theater of Taormina will be all set in the afternoon, the seven microphones have been placed waiting for the absent speakers. After the gladiators, the instrumentalists will come on stage and then the scientists attending the conference will falter into the cypress walk. Only the lizard eventually knows how to raise its head, and, of course, you cannot expect of humans to tow nature. Nature is broken, trapped in its own laws.

Toccata for the Girl with the Burnt Face

Concluding comments

- Most clinical research is either false or not useful
- There are many possible interventions that may improve the efficiency of research practices and make clinical research more credible and more useful
- Empirical meta-research would be useful not only to assess the prevalence of problems, but also to assess the effectiveness and potential harms of interventions that try to improve research

Special thanks

Daniele Fanelli
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