The Crisis in Health Research

Sam Harper, Nicholas B. King

Medicine & Society Elective

2022-04-04

Drinking alcohol key to living past 90

BY
JOE DZIEMIANOWICZ
FOLLOW

NEW YORK DAILY NEWS
Monday, February 19, 2018, 12:07 PM

Cheers to life — seriously.

When it comes to making it into your 90s, booze actually beats exercise, according to a long-term study.

The research, led by University of California neurologist Claudia Kawas, tracked 1,700 nonagenarians enrolled in the 90+ Study that began in 2003 to explore impacts of daily habits on longevity.

Questions for you

1. When you read or hear about new research (publication or talk), what are things that would make you trust the result? What would make you skeptical?

2. Did you ever investigate it further? What did you do, and how did it go?

Outline

What is the replication crisis?

What caused the crisis?

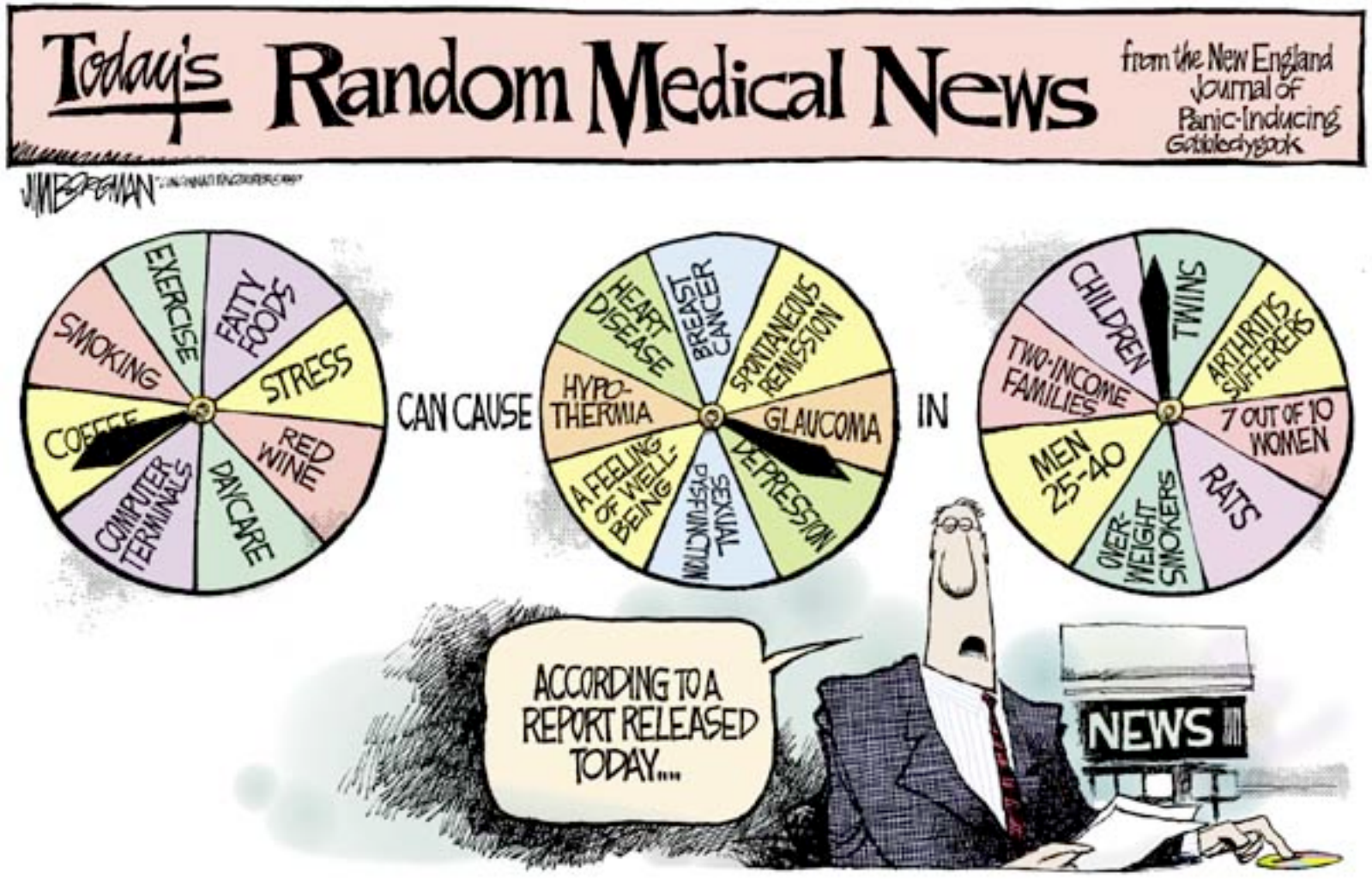
What are some potential solutions?

Outline

What is the replication crisis?

What caused the crisis?

What are some potential solutions?



Source: Borgman (1997)

In theory...

Most studies have low prior probability of being true.

Published research is less likely true if:

- Smaller size study.
- Smaller "true" effects.
- Many tests conducted.
- Flexible study design analysis.
- Conflicts of interest present.
- Sexy/timely/popular topic.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research

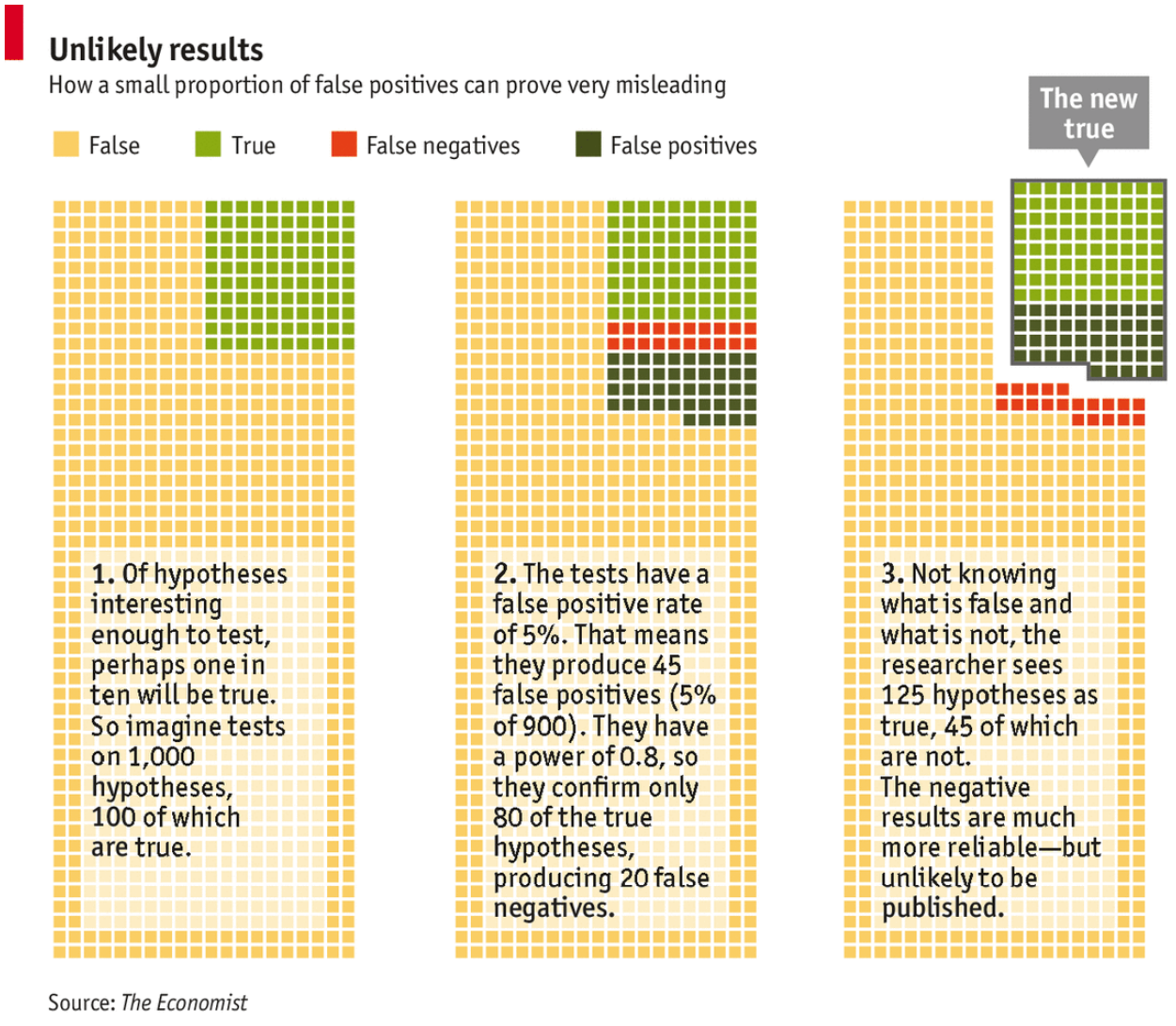
See Ioannidis JPA, 2005

If 10% of hypotheses are true...

AND...we have 80% power...

AND...we only publish 'significant' (p<0.05) findings...

...What is the chance that a published positive finding is actually true?



Source: The Economist

If 10% of hypotheses are true...

AND...we have 80% power...

AND...we only publish 'significant' (p<0.05) findings...

...What is the chance that a published positive finding is actually true?

	"True" Hypothesis Status			
Exp. Result	Positive (null is false)	Negative (null is true)	Total	
"Positive" (reject null)	80	45	125	PPV=64%
"Negative" (accept null)	20	855	875	
Total	100	900	1,000	
	β=80%	α=95%		

Suggests that many published results may be false.

Source: The Economist

2011: Early doubts about ESP in the Psych Lab

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

- Bem (2011) claimed to have found evidence for ESP.
- Two studies that attempted to replicate the Bem study were rejected without peer review at the same journal.

> "This journal does not publish replication studies, whether successful or unsuccessful...I certainly agree that it's desirable that replications are published. The question is where. There are hundreds of journals in psychology...We don't want to be the Journal of Bem Replication."
>
> -JPSP editor Eliot Smith, as told to the New Scientist.

2011: Fraud!

- Psychologist Diederik Stapel admits to fabricating data for many psychology studies.
- 58 papers retracted.
- Concerns raised about reliability of evidence for the entire field of psychology.

SCIENTIFIC ETHICS

Final Report on Stapel Also Blames Field As a Whole

AMSTERDAM—Just when it seemed time to close the book on Diederik Stapel, a new one was opened, quite literally, by Stapel himself. On 28 November, the day on which three investigative panels presented their exhaustive final report on the disgraced social psychologist, Stapel made his first media appearance in more than a year. In a video statement taped at his home by Dutch public television, he said he was deeply sorry and announced he had written an autobiography to explain how it all happened.

The 2:20-minute clip and the appearance of the book the next day caused a new media storm in a country already obsessed with the "Lying Dutchman," as a blogger with *The Washington Post* called him. Many felt that Stapel, fired from Tilburg University in September 2011, was hogging the limelight once again. "Making the announcement on the same day … he's still an ego-tripper," says psychologist Pieter Drenth of the Free University Amsterdam, who chaired one of the committees. Others dispute Stapel's right to earn money off his tale.

But the key message in the joint report—there was one committee for each of the universities where Stapel worked—was a different one: It's not just about Stapel. Colleagues, co-authors, reviewers, and editors at even the most prestigious journals deserve some degree of blame for letting Stapel get away with "blatant" fraud, says the report, which concludes that "the critical function of science has failed on all levels."

The final tally concludes that Stapel made up data in 55 of his 137 papers and in the Ph.D. theses of 10 students he supervised. In another 10 papers, scientific misconduct could not be established beyond a reasonable doubt, either because the original data were missing or Stapel did not confess to fraud, but statistical analyses on the published papers showed in published papers (*Science*, 6 July, p. 21). Simonsohn has raised questions about the work of several social psychologists, including Dirk Smeesters of Erasmus University Rotterdam and Lawrence Sanna of the University of Michigan, Ann Arbor. Amsterdam panel member and statistician Chris Klaassen

Off the tracks. A pirated copy of Stapel's autobiography appeared on the Internet 2 days after it was published.

adapted the method to further reduce chances of a false alarm, Drenth says, adding that the technique "could be used far more broadly when there are suspicions."

In what the report describes as "bycatch," it says that Stapel and his co-authors committed a whole range of smaller sins in papers that weren't outright fraudulent. (It even coins a new Dutch word for sloppy

"I had become addicted to the rhythm of digging, discovering, testing, publishing, scoring, and applause."

—DIEDERIK STAPEL

actual experiments and published papers.

The implications go beyond Stapel, says psycholinguist Willem Levelt, who chaired the Tilburg committee and coordinated the entire investigation. "We're not saying all of social psychology is sloppy science," he says. "But the fact that this could happen shows that the review process has failed from the bottom to the top." Levelt believes the field is taking the message to heart, however. The report praises the "reproducibility project," a large collaborative effort to replicate psychology studies set up by Brian Nosek of the University of Virginia in Charlottesville (*Science*, 30 March, p. 1558), as well as the November issue of *Perspectives on Psychological Science*, which is devoted to analyses of what ails the field and proposals to cure it. "I was impressed by many of the papers," Levelt says. "I have faith that they're cleaning up their shop."

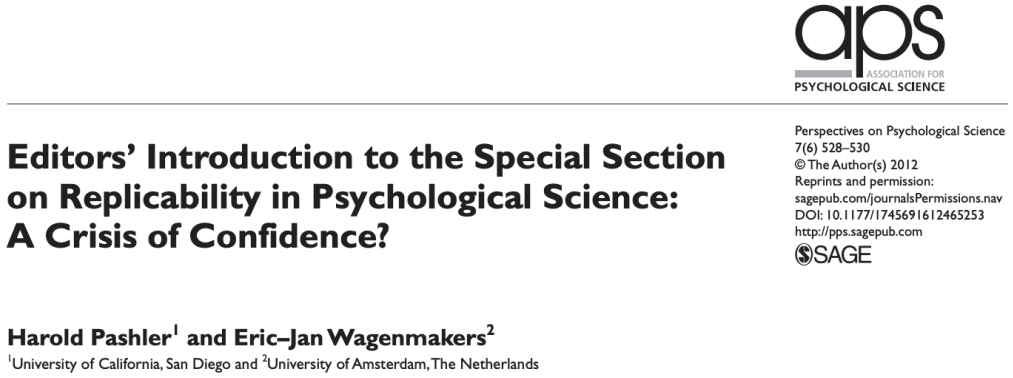
Nosek says the Stapel inquiry is highly unusual in its transparency, the fact that it looked at Stapel's entire corpus of work, and dug deep into the surrounding circumstances. In the case of Harvard University evolutionary biologist Marc Hauser, for instance, a federal investigation found misconduct in six studies, but the university has not released its own report, which reportedly described eight instances of misconduct (*Science*, 14 September, p. 1283). Sanna resigned in May and several of his papers have been retracted, but neither the University of Michigan nor his previous employer, the University of North Carolina, have revealed further information about the case. "We have no idea which of his papers we can trust," Nosek says. "I see the Stapel investigation as a paradigmatic example of how to do it right."

In the wake of the Stapel report, Erasmus University announced that it will take an in-depth look at all of Smeesters's papers, instead of the three that a previous committee studied. Erasmus MC, the university's hospital, will investigate whether the same is possible for Don Poldermans, a cardiologist who was fired in November 2011 after an investigation found "violations of scientific integrity," including fictitious data, in five studies. The

Source: Science, 2012

Systemic concerns

- Fraud and ESP studies
- Psychologists unwilling to share data or code.
- Admission of Questionable Research Practices (QRPs).
- Evidence of manipulating results to get $p<0.05$.
- Well-known results unable to be replicated.



aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?

Harold Pashler[1] **and Eric–Jan Wagenmakers**[2]
[1]University of California, San Diego and [2]University of Amsterdam, The Netherlands

Perspectives on Psychological Science
7(6) 528–530

DOI: 10.1177/1745691612465253
http://pps.sagepub.com
SAGE

Source: Pashler and Wagenmakers 2012

Meanwhile, in biomedical research...

Authors tried to replicate 53 “landmark” drug development studies.

Only 6 (11%) successfully replicated.

Why not?

- Non blinding of experimental vs. control groups.
- Selective reporting of research results.

REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term ‘non-reproduced’ was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.
*Source of citations: Google Scholar, May 2011.

Source: C Glenn Begley and Lee M Ellis, “Drug development: Raise standards for preclinical cancer research”, *Nature*, 483(7391), 2012, pp. 531–3.

Non-replication of candidate gene studies

TABLE 1. Large-scale Efforts to Massively Replicate Reported Candidate-gene Associations[a]

First Author	Disease/Phenotype	Gene Loci Tested	Sample Size (Design)	Replicated Gene Loci[b]
Bosker et al[15]	Major depressive disorder	57	3540 (case-control)	1
Caporaso et al[16]	Smoking (7 phenotypes)	359	4611 (cohort[c])	1
Morgan et al[17]	Acute coronary syndrome	70	1461 (case-control)	0
Richards et al[18]	Osteoporosis (2 phenotypes)	150	19,195 (cohort[d])	3[e]:9[f]
Samani et al[19]	Coronary artery disease	55	4864; 2519 (case-control)	1[g]
Scuteri et al[20]	Obesity (3 phenotypes)	74	6148 (cohort)	0
Sõber et al[21]	Blood pressure	149	1644; 8023 (cohort[h])	0
Wu et al[22]	Childhood asthma	237	1476 (triads[i])	1

[a]The listed studies have been identified through a PubMed search using the strategy (replication [ti] or collaborative [ti] or genome-wide [ti]) and gene* [ti] and (candidate or "previously reported" or "previously proposed") for studies published between 2007 and 2010 (last search 29 July 2010) and complemented with eligible studies from those analyzed in the paper by Siontis et al.[24]

[b]With proper, stringent control for multiple comparisons.

[c]Two cohorts with combined analysis.

[d]Five cohorts with combined analysis.

[e]Femoral bone density.

[f]Spinal bone density; the 3 gene loci associated with femoral bone density were also associated with femoral bone density.

[g]2 case-control studies analyzed separately; the gene locus significantly associated in the smaller study also had the strongest association in the larger study.

[h]Initial discovery cohort of size 1644, with subsequent replication in 2 cohorts of size 1830 and 1823 and one case-control study with 2401 cases and 1969 controls.

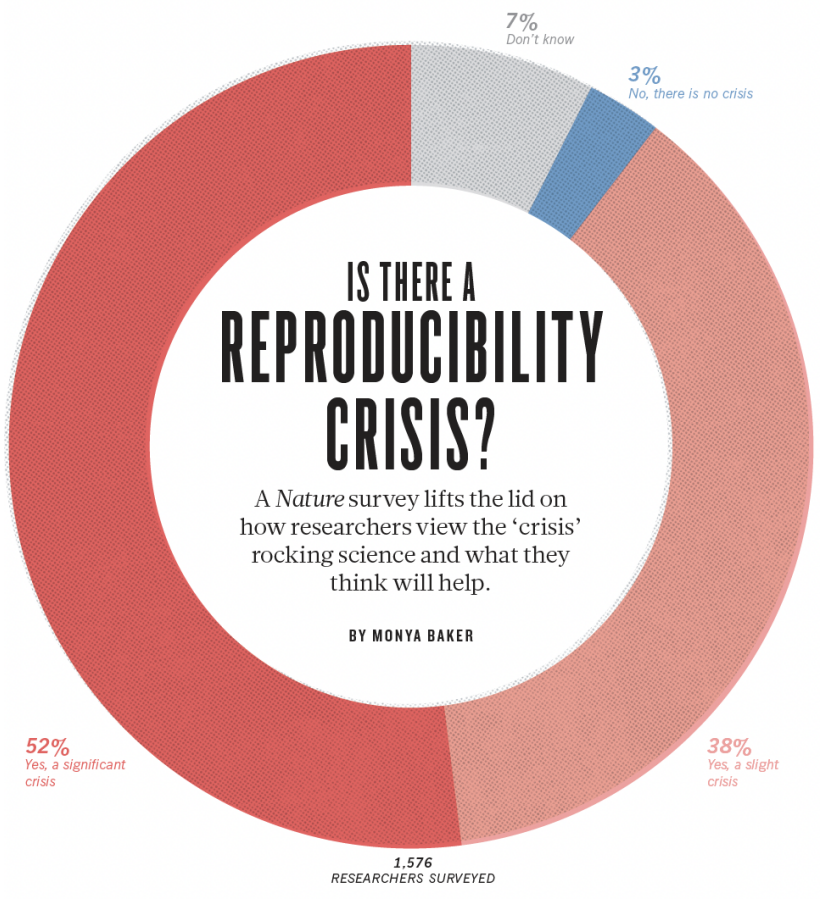
[i]Case-parent triad design, with 492 triads consisting of an asthmatic child and both parents.

Ioannidis (2011)

Et cetera, et cetera...

Also evidence of high-profile non-replication studies in:

- Psychiatry
- Neurobiology
- Chemistry
- Biology
- Ecology & evolution
- Oceanography
- Likely any place we have not looked...



See Ritchie, *Science Fictions* (2019), Baker *Nature* (2018)

Concerns about reliablility of findings across multiple disciplines led to collaborative, large-scale *replication* studies.

Distinctions between commonly used terms

Replication

Using using independent investigators, methods, data, equipment, and protocols, we arrive at the same conclusions and/or the same estimate of the effect.

There can be good reasons why findings do not replicate.

Reproducibility

If we start from the *same* data gathered by the scientist we can reproduce the same results, p-values, confidence intervals, tables and figures as in the original report.

There are fewer reasons for non-reproducibility.

National Academy of Sciences (2019)

Definitions

- May not be used consistently.
- Clarify whether data and analysis are different from original study.

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Source: https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

Method

Starting in November 2011, we constructed a protocol for selecting and conducting high-quality replications (*24*). Collaborators joined the project, selected a study for replication from the available studies in the sampling frame, and were guided through the replication protocol. The replication protocol articulated the process of selecting the study and key effect from the available articles, contacting the original authors for study materials, preparing a study protocol and analysis plan, obtaining review of the protocol by the original authors and other members within the present project, registering the protocol publicly, conducting the replication, writing the final report, and auditing the process and analysis for quality control. Project coordinators

Nosek et al., 2015. Also see letters and responses: http://bit.ly/1Xg9lQl

RESEARCH ARTICLE

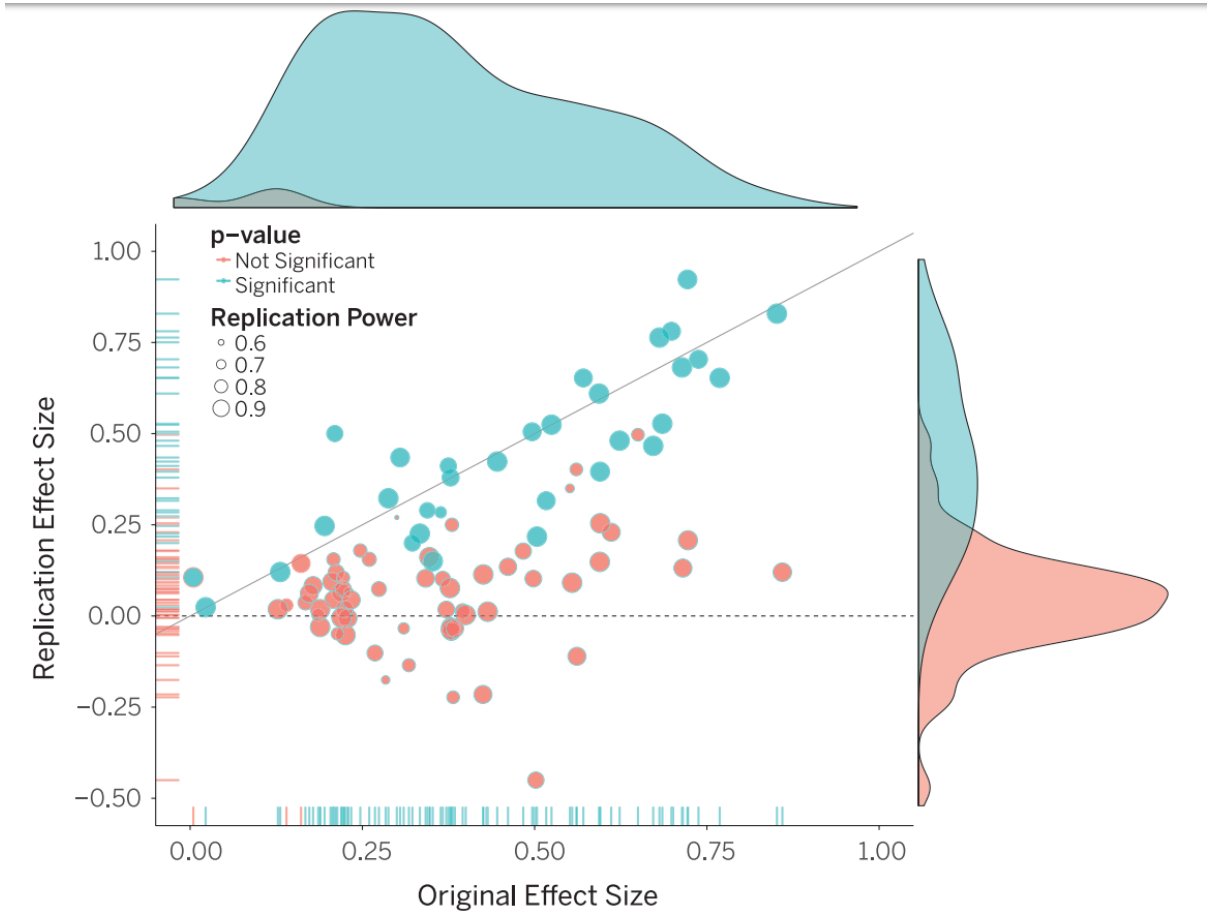
PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Effect sizes are much lower in replication studies.



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Nosek et al. (2015)

No better in Economics

- Increased the number of participants by a factor of five
- Preregistered their study and analysis designs before any data collection.

ECONOMICS

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

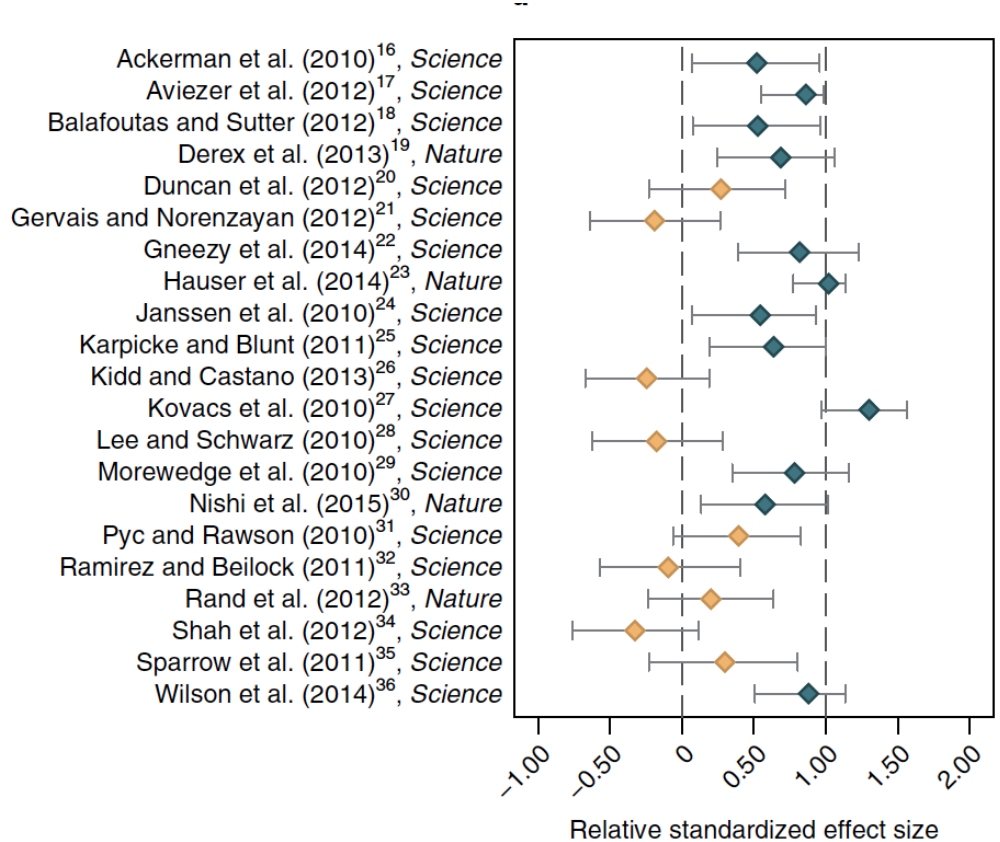
The replicability of some scientific findings has recently been called into question. To contribute data about replicability in economics, we replicated 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014. All of these replications followed predefined analysis plans that were made publicly available beforehand, and they all have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original. The replicability rate varies between 67% and 78% for four additional replicability indicators, including a prediction market measure of peer beliefs.

Camerer et al. (2016)

Surely the "top" journals are better, right?

"We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size"

"The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility."



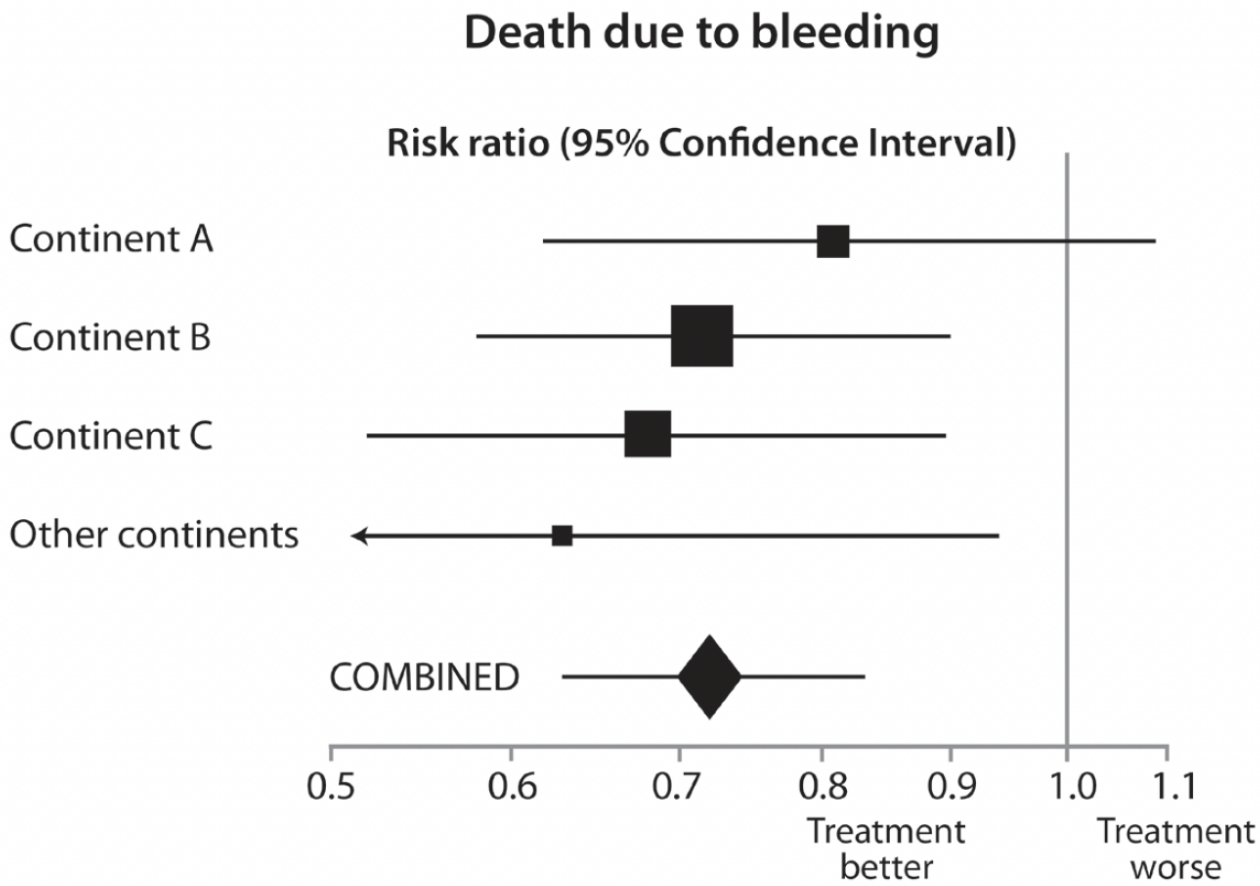
Camerer et al. (2018)

What is a replication?

- *Exact* replication is impossible (without time travel)
- New studies involve new subjects, different environment.
- Should not expect exactly the same results.
- Better to think of it as a study that can tell us something about the veracity of a prior claim.

See Nosek and Errington, PLoS Biology 2020. GIF from tenor.com.

Can think of cluster-randomized trials as replications



Effects of tranexamic acid on death among trauma patients with significant haemorrhage, overall and by continent of participants (unpublished data from CRASH-2: *Lancet* 2010;376:23-32).

Source: Evans et al., *Testing Treatments* (2011)

Do we really mean 'replication'?

VIEWPOINT

Christopher P. Childers, MD, PhD
Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, California.

Melinda Maggard-Gibbons, MD, MSHS
Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, California.

Replication Studies for Database Research

Many clinical questions cannot be answered with a randomized trial because of issues surrounding ethics, cost, and practicality. Observational studies can help fill this void; however, the ability to translate findings into clinical practice depends on the quality and rigor of study design and statistical analysis. Over the past several years, *JAMA Surgery* has focused on this concept by publishing guidelines to improve the quality of research performed using large databases.[1] Recently, we proposed the concept of "replication studies."[2] The idea is simple—external researchers reproduce a study's results and perform novel sensitivity analyses to ascertain how consistent the findings are, the purpose of which is to confirm that the sign and magnitude of the primary coefficient are accurate. Signals that persist are more "robust" and are therefore more likely to have the desired outcome in clinical practice. Journals would conclusions about how robust they perceive the original findings.

How Would Journals Manage Replication Studies?

Journals and editors have several options. The most straightforward would be to solicit replication studies after an article is published, in much the same way journals allow for Letters to the Editor. This would require extending the time frame for such submissions—perhaps to 3 months after the index publication. If the results support the existing study, these could be published alone. If the results conflict, we suggest allowing the original authors the opportunity to respond and soliciting a piece from a third-party to arbitrate the conflict. If journals choose to fully embrace replication studies, an alternative—and likely more impactful—way to publish these studies would be for journals to

Do we really mean 'replication'?

VIEWPOINT

Christopher P. Childers, MD, PhD
Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, California.

Melinda Maggard-Gibbons, MD, MSHS
Department of Surgery, David Geffen School of Medicine at UCLA, Los Angeles, California.

~~Replication~~ Studies for Database Research

ROBUSTNESS

Many clinical questions cannot be answered with a randomized trial because of issues surrounding ethics, cost, and practicality. Observational studies can help fill this void; however, the ability to translate findings into clinical practice depends on the quality and rigor of study design and statistical analysis. Over the past several years, *JAMA Surgery* has focused on this concept by publishing guidelines to improve the quality of research performed using large databases.[1] Recently, we proposed the concept of "replication studies."[2] The idea is simple—external researchers reproduce a study's results and perform novel sensitivity analyses to ascertain how consistent the findings are, the purpose of which is to confirm that the sign and magnitude of the primary coefficient are accurate. Signals that persist are more "robust" and are therefore more likely to have the desired outcome in clinical practice. Journals would conclusions about how robust they perceive the original findings.

How Would Journals Manage Replication Studies?

Journals and editors have several options. The most straightforward would be to solicit replication studies after an article is published, in much the same way journals allow for Letters to the Editor. This would require extending the time frame for such submissions—perhaps to 3 months after the index publication. If the results support the existing study, these could be published alone. If the results conflict, we suggest allowing the original authors the opportunity to respond and soliciting a piece from a third-party to arbitrate the conflict. If journals choose to fully embrace replication studies, an alternative—and likely more impactful—way to publish these studies would be for journals to

Considering 'robustness'

- What if we tweak prior methods?
- Many decisions affect the 'final' result.
- Would it matter?

Reanalyses of Randomized Clinical Trial Data

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

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

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

Shanil Ebrahim, et al., "Reanalyses of randomized clinical trial data", *JAMA*, 312(10), 2014, pp. 1024–32.

What leads to different answers?

- Reducing measurement error:

Table 3. Reanalyses Producing a Change in Whom to Treat

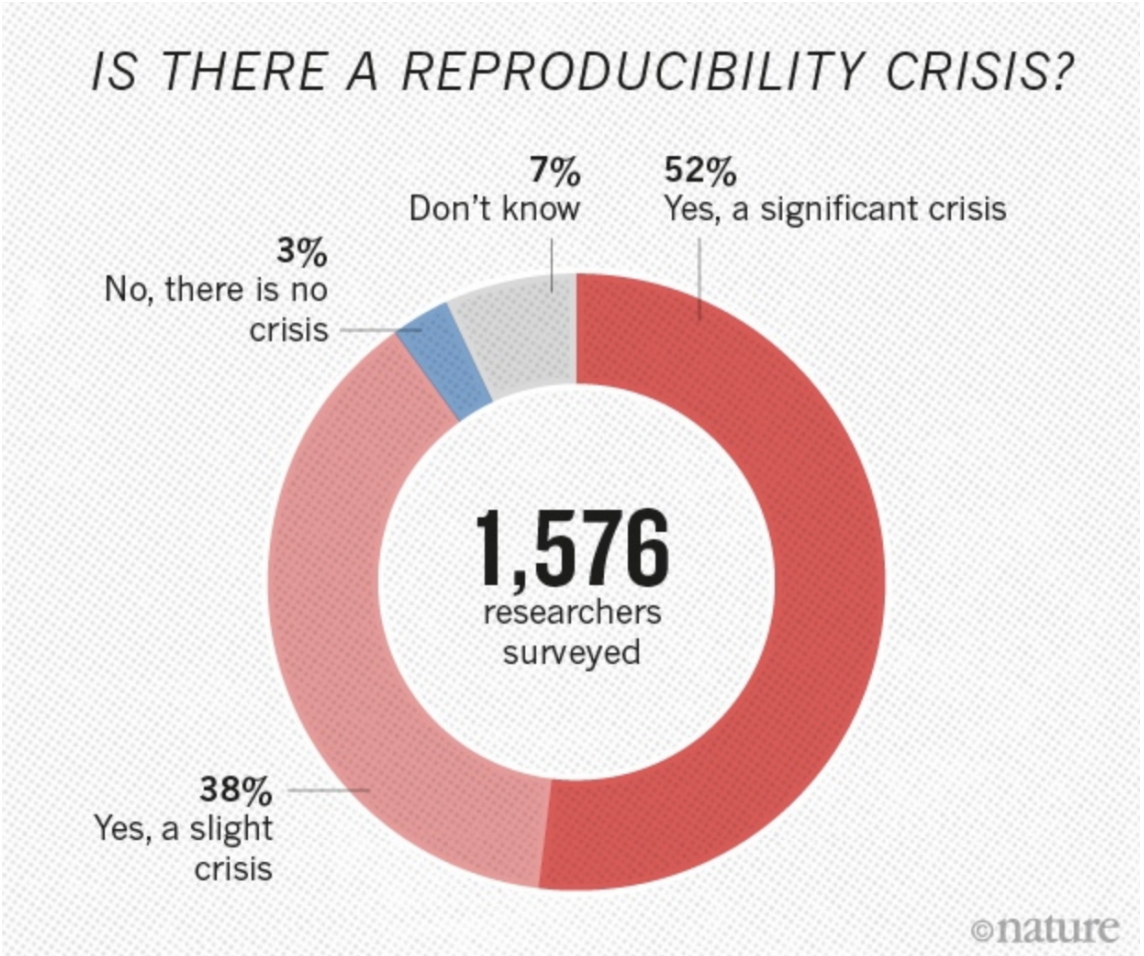
Source	Patient Population	Intervention Comparators	Primary Outcome	Original Trial Interpretation	Differences in Methods Used in the Reanalysis	Change in Finding	Change in Interpretation
Johnston et al,[27] 1985	Coronary artery disease with regional left ventricular dysfunction	Pindolol vs propranolol	Left ventricular ejection fraction at rest and exercise	No difference between pindolol and propranolol	Regional wall motion abnormalities reanalyzed with a computer-assisted rather than visual method	Pindolol superior to propranolol	Treat with pindolol rather than propranolol (more patients to be treated with the newer treatment)

- Allowing for effect modification:

Brooks et al,[17] 1998	Alzheimer disease	Acetyl L-carnitine vs placebo	Performance on the cognitive subscale of the Alzheimer Disease Assessment Scale	No difference between acetyl L-carnitine and placebo	Cognitive subscale reanalyzed as rate of change; analysis included test for interaction between drug effect and age	Test of drug × age interaction statistically significant, with younger patients benefiting more from treatment than older patients	Treat younger patients (different patients)

Scentists think there is a "reproducibility" crisis

or a "slight" crisis? 🤔



Baker *Nature* (2018)

Outline

What is the replication crisis?

What caused the crisis?

What are some potential solutions?

Potential sources of "bias" in published research

Usual explanations

Confounding, measurement error,
selection bias, model misspecification, etc.

Problems with integrity

- Fraud/data manipulation/fabrication.
- Poor design / inadequate power.
- NHST: Publication bias.
- NHST: P-hacking.
- Financial ties/ideological commitments.
- Careerism.
- Lack of transparency.

Table 1 Some prominent cases of data fraud in clinical trials

Name	Allegations/findings	Outcome	Key references
Roger Poisson	Falsification of eligibility data on multi-center breast cancer trials	Barred from research funding (8 years)	Fisher and Redmond [3], Weir and Murray [4]
Werner Bezwoda	Fabrication and falsification of data on single institution breast cancer trials	Dismissed from position	Horton [5], Weiss et al. [6]
Robert Fiddes	Fabrication and falsification of data and entering ineligible patients on multi-center industry-supported clinical trials	Prison sentence (15 months)	Eichenwald and Kolata [7], Swaminathan and Avery [8]
Harry Snyder Renee Peugeot	Falsification of data on single-institution clinical trials for a biotech firm	Prison sentences (3 years; 2.5 years), financial restitution	Birch and Cohen [9], Grant [10]
Yoshiaka Fujii	Fabrication of data on clinical trials in post-operative nausea and vomiting	Dismissed from position, 183 papers retracted	Kranke et al. [14], Carlisle [11]
Anil Potti	Falsification of genomics data used in predictive modelling for cancer clinical trials	Resigned position, 11 papers retracted	Baggerly and Coombes [17]
Hiroaki Matsubara	Fabrication and falsification of data on clinical trials of antihypertensive agent valsartan	Resigned position, 9 papers retracted	Husten [19], Oransky [18]

The coronavirus pandemic has not helped.

- High profile studies of 96k patients across 671 hospitals.
- Claimed hydroxychloroquine increased mortality from COVID-19.
- Immediately led to WHO halting the hydroxychloroquine arm of its global trials.
- Researchers subsequently questioned irregularities in the data, ethics review, protocols, statistical analysis.
- Two papers (other in *NEJM*) were retracted.

See reporting by Davey et al. in *The Guardian*

Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Mandeep R Mehra, Sapan S Desai, Frank Ruschitzka, Amit N Patel

Summary

Background Hydroxychloroquine or chloroquine, often in combination with a second-generation m… ide, are b… widely used for treatment of COVID-19, despite no conclusive evidence of their benefit. Althoug… nerall… fe whe… used for approved indications such as autoimmune disease or malaria, the safety and bene… f th… reatment regimens are poorly evaluated in COVID-19.

Methods We did a multinational registry analysis of the use of hydroxychloroquine … oquine with… thout a macrolide for treatment of COVID-19. The registry comprised data from 671 hosp… s in s… ntinents. We included patients hospitalised between Dec 20, 2019, and April 14, 2020, with a positive laboratory t… g for SARS-CoV-2. Patients who received one of the treatments of interest within 48 h of diagn… included in … f four treatment groups (chloroquine alone, chloroquine with a macrolide, hydroxychlor… ine alone, or hydroxychloroquine with a macrolide), and patients who received none of these treatments formed … control gr… . Patients for whom one of the treatments of interest was initiated more than 48 h after diagnosis or … ile they w… on mechanical ventilation, as well as patients who received remdesivir, were excluded. The main outc… of int… st were in-hospital mortality and the occurrence of de-novo ventricular arrhythmias (… tained o… ed ventricular tachycardia or ventricular fibrillation).

Findings 96 032 patients (mean age 53·8 years, 46·3% women) … OVID-19 were hospitalised during the study period and met the inclusion criteria. Of the… patie… were in the treatment groups (1868 received chloroquine, 3783 received chloroquine with … macro… e, 3016 … eived hydroxychloroquine, and 6221 received hydroxychloroquine with a macrolide) and … pati… in the control group. 10 698 (11·1%) patients died in hospital. After controlling for multiple … fou… ctors (… sex, race or ethnicity, body-mass index, underlying cardiovascular disease and its risk fac… diabetes… derlying lung disease, smoking, immunosuppressed condition, and baseline disease severity), w… mpared wi… ortality in the control group (9·3%), hydroxychloroquine (18·0%; hazard ratio 1·335, 95%… 1·22… 457), hydr… ychloroquine with a macrolide (23·8%; 1·447, 1·368–1·531), chloroquine (16·4%; 1·365, 1·…18–1·531), … chloroquine with a macrolide (22·2%; 1·368, 1·273–1·469) were each independently associated … an increased r… of in-hospital mortality. Compared with the control group (0·3%), hydroxychloroquine (6… ; 2·36… ·935–2·900), hydroxychloroquine with a macrolide (8·1%; 5·106, 4·106–5·983), chloroquine (4·3%; … 1, 2… 0–4·596), and chloroquine with a macrolide (6·5%; 4·011, 3·344–4·812) were independently associate… an inc… ed risk of de-novo ventricular arrhythmia during hospitalisation.

Interpretati… We v… unable… nfirm a benefit of hydroxychloroquine or chloroquine, when used alone or with a macro… on in… spital outcomes for COVID-19. Each of these drug regimens was associated with decreased in-hospita… i… reased frequency of ventricular arrhythmias when used for treatment of COVID-19.

Funding William… vey Distinguished Chair in Advanced Cardiovascular Medicine at Brigham and Women's Hospital.

P… ed Online … 22, 2020 https://doi.org/10.1016/S0140-6736(20)31180-6

This online publication has been corrected. The corrected version first appeared at thelancet.com on May 29, 2020

See Online/Comment https://doi.org/10.1016/S0140-6736(20)31174-0

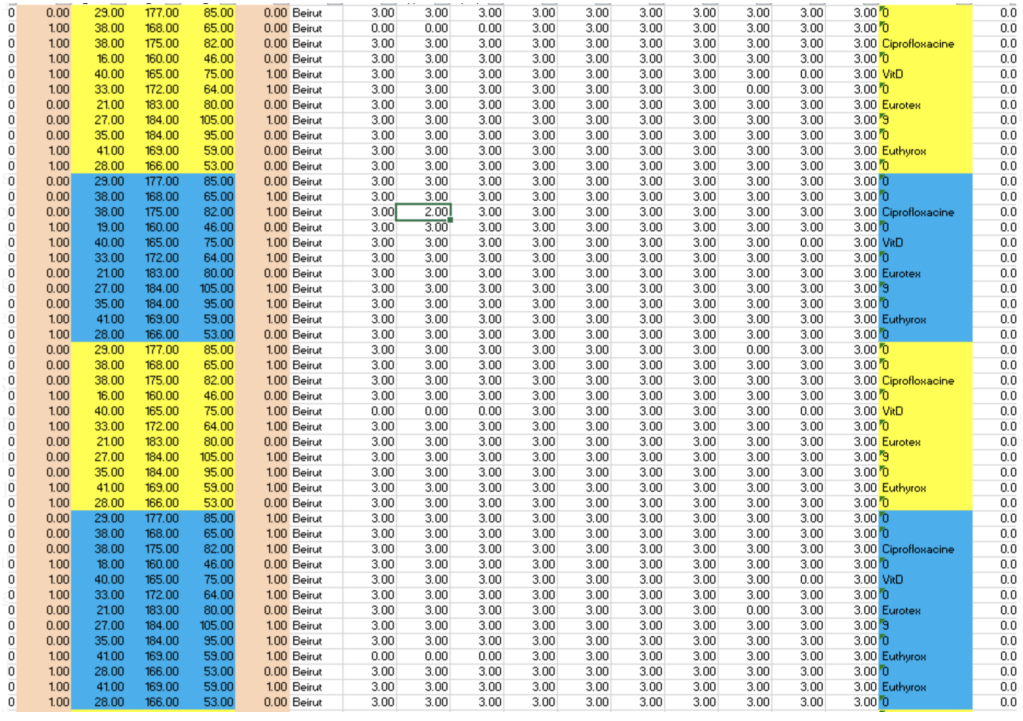
Brigham and Women's Hospital Heart and Vascular Center and Harvard Medical School, Boston, MA, USA (Prof M R Mehra MD); Surgisphere Corporation, Chicago, IL, USA (S S Desai MD); University Heart Center, University Hospital Zurich, Zurich, Switzerland (Prof F Ruschitzka MD); Department of Biomedical Engineering, University of Utah, Salt Lake City, UT, USA (A N Patel MD); and HCA Research Institute, Nashville, TN, USA (A N Patel)

Correspondence to: Prof Mandeep R Mehra, Brigham and Women's Hospital Heart and Vascular Center and Harvard Medical School, Boston, MA 02115, USA mmehra@bwh.harvard.edu

RETRACTED

The coronavirus pandemic has not helped.

- Ivermectin known to be effective for parasitic diseases.
- Lebanese trial claimed *huge effect* of ivermectin on lowering SARS-CoV2 viral load.
- Requests for trial data to verify the results by independent researchers were denied.
- Researchers subsequently gained access, found serious irregularities.
- Ultimately retracted.



0	0.00	29.00	177.00	85.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0		0.0
0	1.00	38.00	168.00	65.00	0.00	Beirut	0.00	0.00	0.00	3.00	3.00	3.00	3.00	3.00	3.00	0		0.0
0	1.00	38.00	175.00	82.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Ciprofloxacine		0.0
0	1.00	16.00	160.00	46.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0		0.0
0	1.00	40.00	165.00	75.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	VitD	0.0
0	1.00	33.00	172.00	64.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	3.00	0	0.0
0	0.00	21.00	183.00	80.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Eurotex	0.0
0	0.00	27.00	184.00	105.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	9	0.0
0	0.00	35.00	184.00	95.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	41.00	169.00	59.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Euthyrox	0.0
0	1.00	28.00	166.00	53.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	29.00	177.00	85.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	38.00	168.00	65.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	38.00	175.00	82.00	1.00	Beirut	3.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Ciprofloxacine	0.0
0	1.00	19.00	160.00	46.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	40.00	165.00	75.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	VitD	0.0
0	1.00	33.00	172.00	64.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	21.00	183.00	80.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Eurotex	0.0
0	0.00	27.00	184.00	105.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	9	0.0
0	0.00	35.00	184.00	95.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	41.00	169.00	59.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Euthyrox	0.0
0	1.00	28.00	166.00	53.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	29.00	177.00	85.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	3.00	0	0.0
0	0.00	38.00	168.00	65.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	38.00	175.00	82.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Ciprofloxacine	0.0
0	1.00	16.00	160.00	46.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	40.00	165.00	75.00	1.00	Beirut	0.00	0.00	0.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	VitD	0.0
0	1.00	33.00	172.00	64.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	21.00	183.00	80.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Eurotex	0.0
0	0.00	27.00	184.00	105.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	9	0.0
0	0.00	35.00	184.00	95.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	41.00	169.00	59.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Euthyrox	0.0
0	1.00	28.00	166.00	53.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	29.00	177.00	85.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	38.00	168.00	65.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	38.00	175.00	82.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Ciprofloxacine	0.0
0	1.00	18.00	160.00	46.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	40.00	165.00	75.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	VitD	0.0
0	1.00	33.00	172.00	64.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	0.00	21.00	183.00	80.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00	3.00	3.00	Eurotex	0.0
0	0.00	27.00	184.00	105.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	9	0.0
0	0.00	35.00	184.00	95.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	41.00	169.00	59.00	1.00	Beirut	0.00	0.00	0.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Euthyrox	0.0
0	1.00	28.00	166.00	53.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0
0	1.00	41.00	169.00	59.00	1.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	Euthyrox	0.0
0	1.00	28.00	166.00	53.00	0.00	Beirut	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0	0.0

See reporting by Meyerowitz-Katz

ORIGINAL ARTICLE

Effect of Early Treatment with Ivermectin among Patients with Covid-19

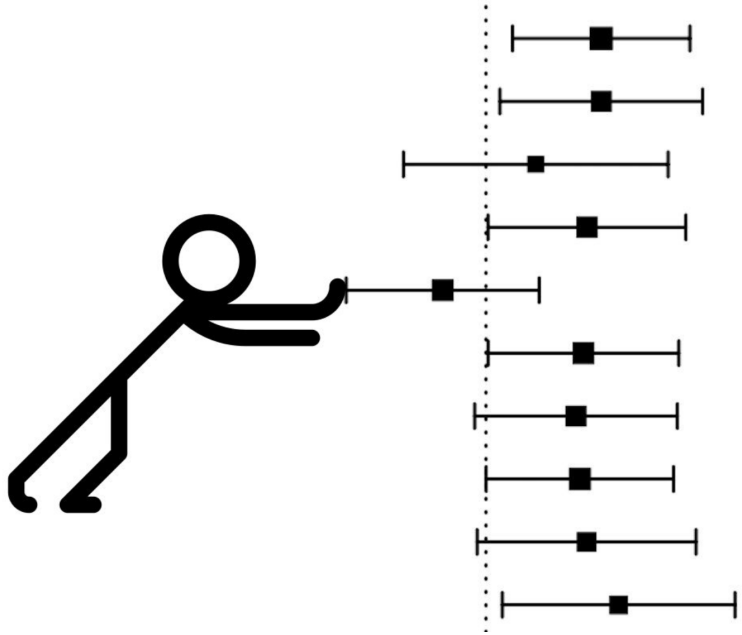
Gilmar Reis, M.D., Ph.D., Eduardo A.S.M. Silva, M.D., Ph.D., Daniela C.M. Silva, M.D., Ph.D., Lehana Thabane, Ph.D., Aline C. Milagres, R.N., Thiago S. Ferreira, M.D., Castilho V.Q. dos Santos, Vitoria H.S. Campos, Ana M.R. Nogueira, M.D., Ana P.F.G. de Almeida, M.D., Eduardo D. Callegari, M.D., Adhemar D.F. Neto, M.D., Ph.D., et al., for the TOGETHER Investigators*

March 30, 2022
DOI: 10.1056/NEJMoa2115869

Table 2. Effect of Ivermectin as Compared with Placebo on Covid-19–Related Hospitalization or Extended Observation in an Emergency Setting.*

Population and Trial Group	Population Size	Patients with Primary-Outcome Event	Relative Risk (95% Bayesian Credible Interval)
	no.	*no. (%)*	
Intention-to-treat population			
Ivermectin	679	100 (14.7)	0.90 (0.70–1.16)
Placebo	679	111 (16.3)	Reference
All	1358	211 (15.5)	—

A lot of irreproducible or unreliable research stems from Null Hypothesis Significance Testing (NHST).

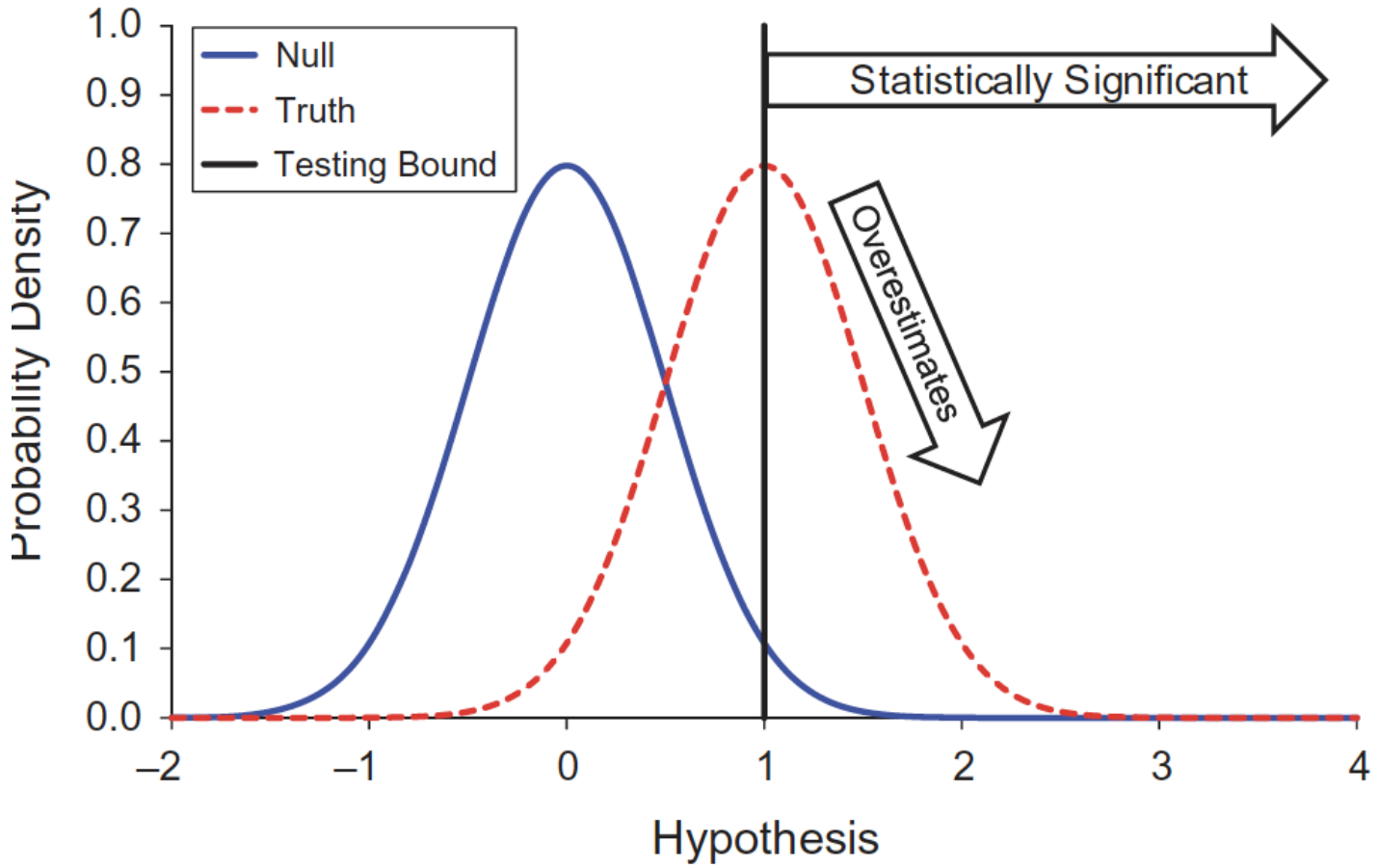


https://mobile.twitter.com/wviechtb/status/1228327958810648576/photo/1

NHST facilitates non-replication

Study results are sampled from the (---) distribution, but we only see 'statistically significant' ones

Implications for planning, meta-analysis, and replication studies.

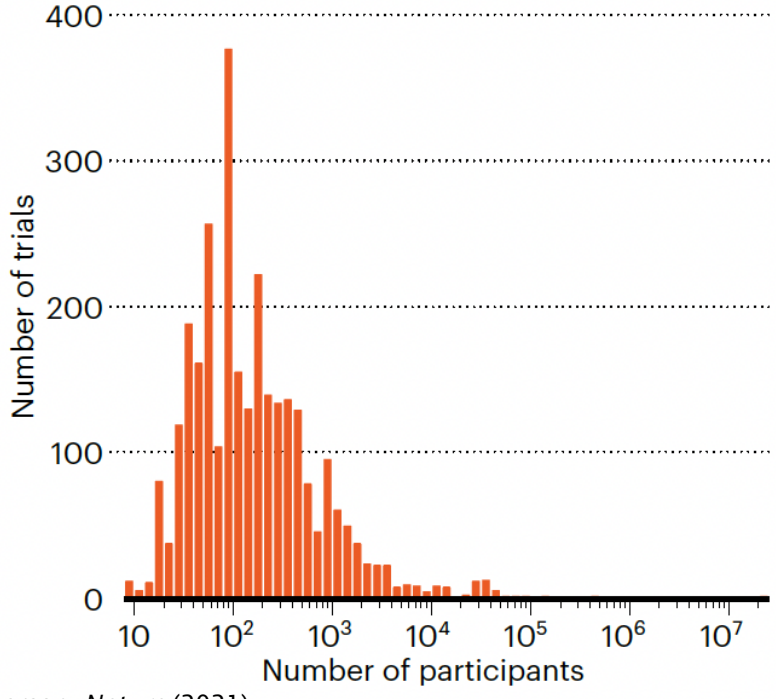


Lash (2017)

The coronavirus pandemic has not helped.

SMALL SAMPLES

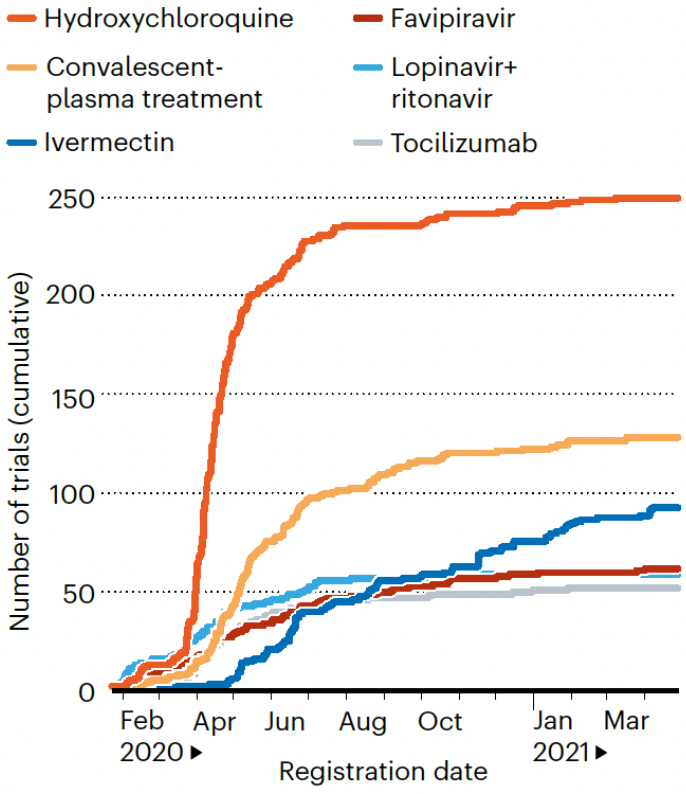
In one database of COVID-19 trials, 40% stated that they were enrolling fewer than 100 patients — a sample size that is generally too small to be useful.



Source: Pearson, *Nature* (2021)

TOO MANY TRIALS?

Studies assessing drugs against COVID-19 included 250 trials of hydroxychloroquine — a duplication that researchers say represents wasted effort.



Researcher "degrees of freedom" are difficult to control

How are analyses conducted?

- collect the data over many months.
- finish recording and merging.
- run *one* regression.
- new regression, different controls.
- now a different functional form.
- new regression, different measures.
- yet another regression on subset.
- have 100 or 1000 estimates.
- 1 or maybe 5 results in the paper.

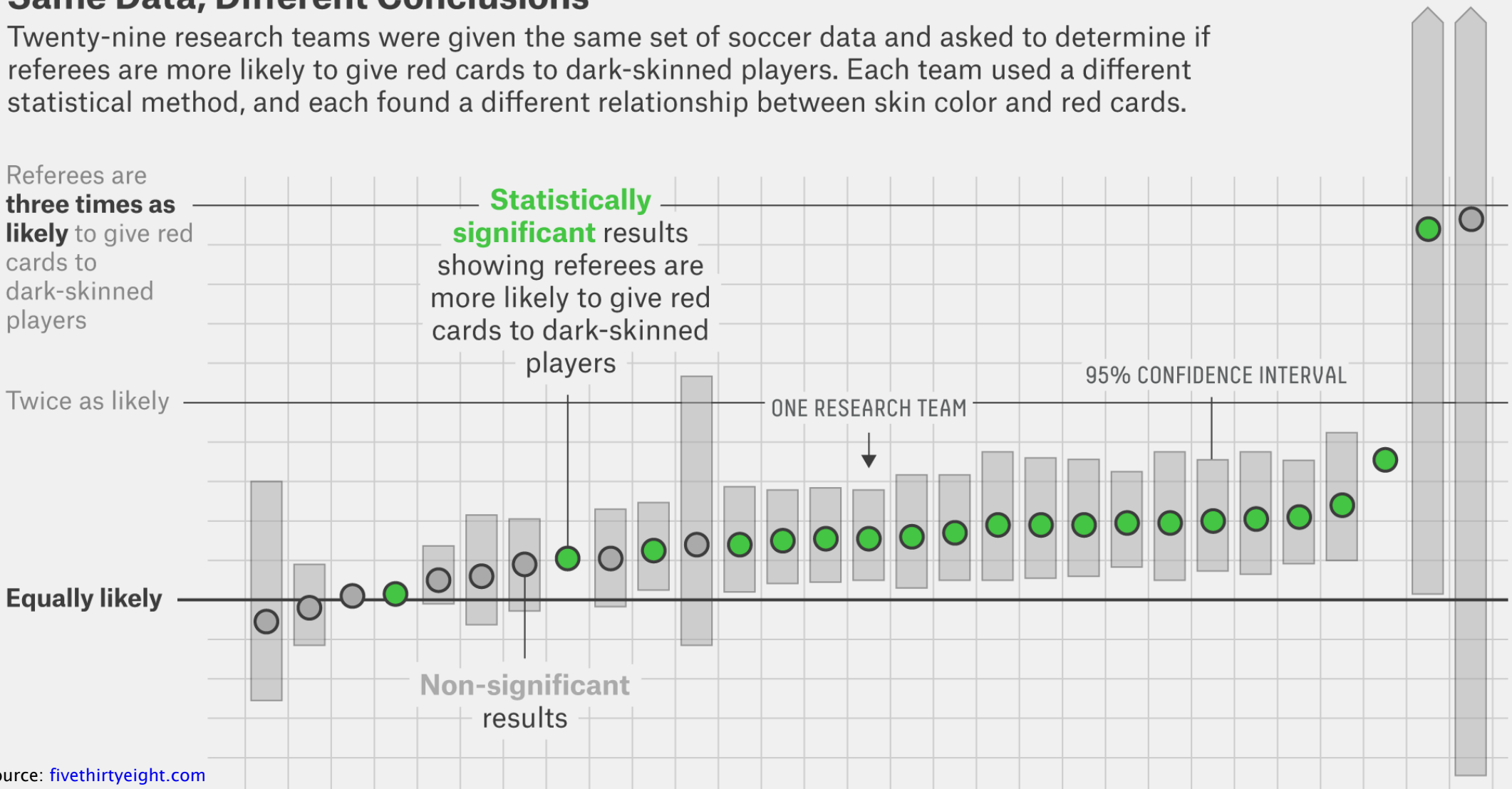
What's the problem?

- Some result is designated as the “correct” one, only *after* looking at the estimates.
- Is this a true test of a hypothesis or just confirmation bias?
- This is "p-hacking"

Source: Gary King

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



Source: fivethirtyeight.com

FIVETHIRTYEIGHT

SOURCE: BRIAN NOSEK ET AL.

Let's do some hacking!

Go to https://projects.fivethirtyeight.com/p-hacking/ and answer this question:

Will the 2022 US midterm elections affect the economy?

03:00

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

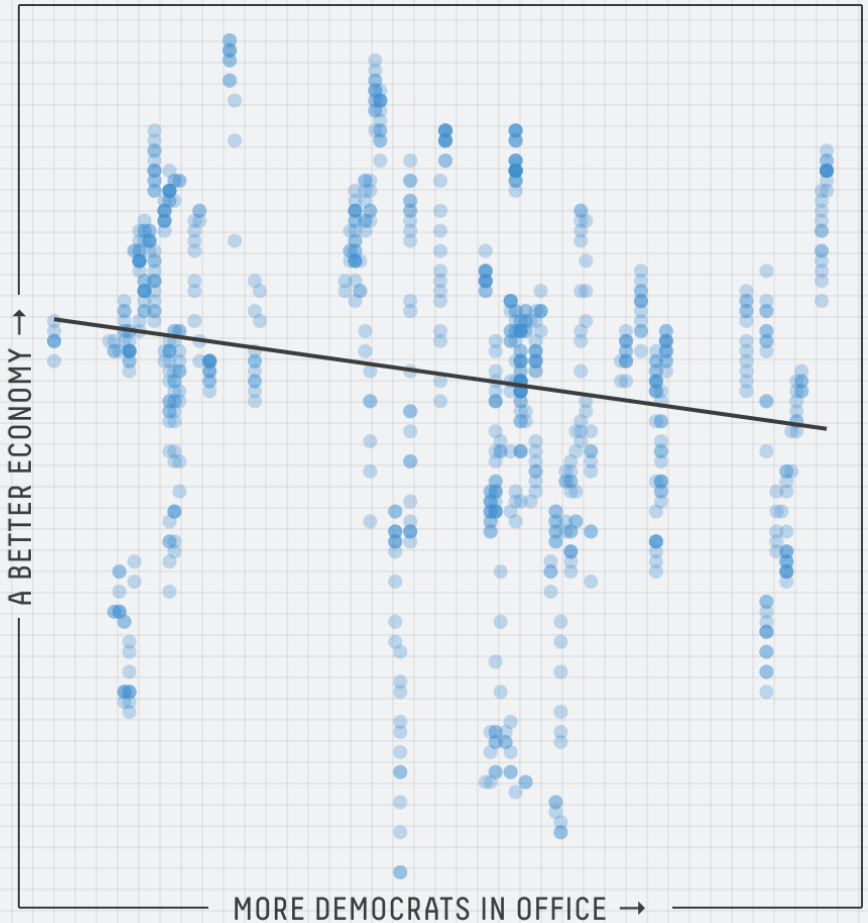
- [x] **Employment**
- [ ] **Inflation**
- [ ] **GDP**
- [ ] **Stock prices**

Other options

- [ ] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

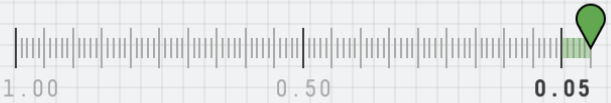
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **negative** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

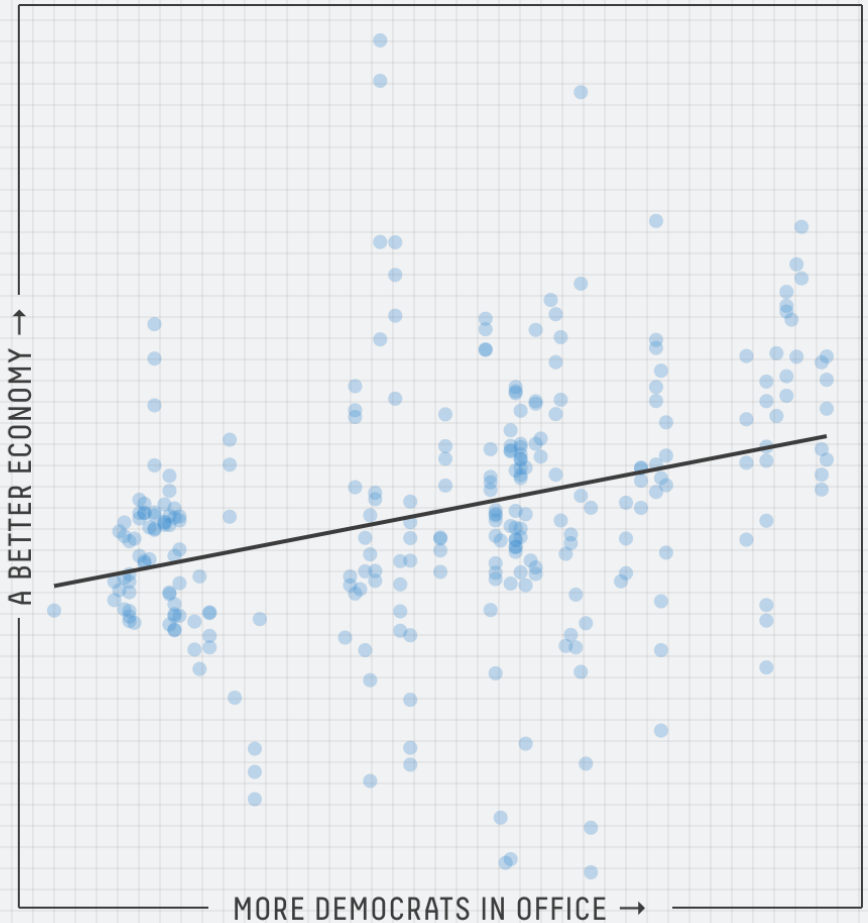
- [ ] **Employment**
- [ ] **Inflation**
- [x] **GDP**
- [ ] **Stock prices**

Other options

- [ ] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

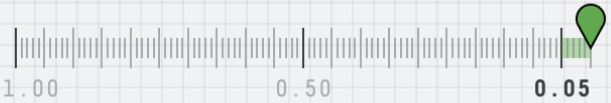
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **positive** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

How do we know there is p-hacking?

(1) Look at what people are doing.

Two estimates:

- HR=0.90, 95%CI: 0.81, 0.99 "Significantly lower"
- HR=0.89, 95%CI: 0.78, 1.00009 "No difference"

Normalization of Testosterone Levels After Testosterone Replacement Therapy Is Associated With Decreased Incidence of Atrial Fibrillation

Rishi Sharma, MD, MHSA; Olurinde A. Oni, MBBS, MPH; Kamal Gupta, MD; Mukut Sharma, PhD; Ram Sharma, PhD; Vikas Singh, MD, MHSA; Deepak Parashara, MD; Surineni Kamalakar, MBBS, MPH; Buddhadeb Dawn, MD; Guoqing Chen, MD, PhD, MPH; John A. Ambrose, MD; Rajat S. Barua, MD, PhD

Background—Atrial fibrillation (AF) is the most common cardiac dysrhythmia associated with significant morbidity and mortality. Several small studies have reported that low serum total testosterone (TT) levels were associated with a higher incidence of AF. In contrast, it is also reported that anabolic steroid use is associated with an increase in the risk of AF. To date, no study has explored the effect of testosterone normalization on new incidence of AF after testosterone replacement therapy (TRT) in patients with low testosterone.

Methods and Results—Using data from the Veterans Administrations Corporate Data Warehouse, we identified a national cohort of 76 639 veterans with low TT levels and divided them into 3 groups. Group 1 had TRT resulting in normalization of TT levels (normalized TRT), group 2 had TRT without normalization of TT levels (nonnormalized TRT), and group 3 did not receive TRT (no TRT). Propensity score–weighted stabilized inverse probability of treatment weighting Cox proportional hazard methods were used for analysis of the data from these groups to determine the association between post-TRT levels of TT and the incidence of AF. Group 1 (40 856 patients, median age 66 years) had significantly lower risk of AF than group 2 (23 939 patients, median age 65 years; hazard ratio 0.90, 95% CI 0.81–0.99, P=0.0255) and group 3 (11 853 patients, median age 67 years; hazard ratio 0.79, 95% CI 0.70–0.89, P=0.0001). There was no statistical difference between groups 2 and 3 (hazard ratio 0.89, 95% CI 0.78–1.0009, P=0.0675) in incidence of AF.

Conclusions—These novel results suggest that normalization of TT levels after TRT is associated with a significant decrease in the incidence of AF. (***J Am Heart Assoc.* 2017;6:e004880. DOI: 10.1161/JAHA.116.004880.**)

Key Words: atrial fibrillation • testosterone • testosterone replacement therapy

https://www.ahajournals.org/doi/abs/10.1161/jaha.116.004880

Good advice on how to get $p<0.05$

Here's some things to do.

First, look to see if there are weird outliers (in terms of how much they ate). If there seems to be a reason they are different, pull them out but specially note why you did so, so that this can be described in the method.

Third, look at a bunch of different DVs. These might include

- # pieces of pizza
- # trips
- Fill level of plate
- Did they get dessert
- Did they order a drink
- and so on . . .

Second, think of all the different ways you can cut the data and analyze subsets of it to see when this relationship holds. For instance, if it works on men but not women, we have a moderator. Here are some groups you'll want to break out separately:

- Males
- Females
- Lunch goers
- Dinner goers
- People sitting alone
- People eating with groups of 2
- People eating in groups of 2+
- People who order alcohol
- People who order soft drinks
- People who sit close to buffet
- People who sit far away
- and so on . . .

"I gave her a data set of a self-funded, failed study which had null results... I said, 'This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set.' I had three ideas for potential Plan B, C, & D directions (since Plan A had failed)." -blog, 2016

Enterprising grad students found:

- impossible values
- incorrect ANOVA results
- dubious p-values

Wansink denied requests for access to the original data.

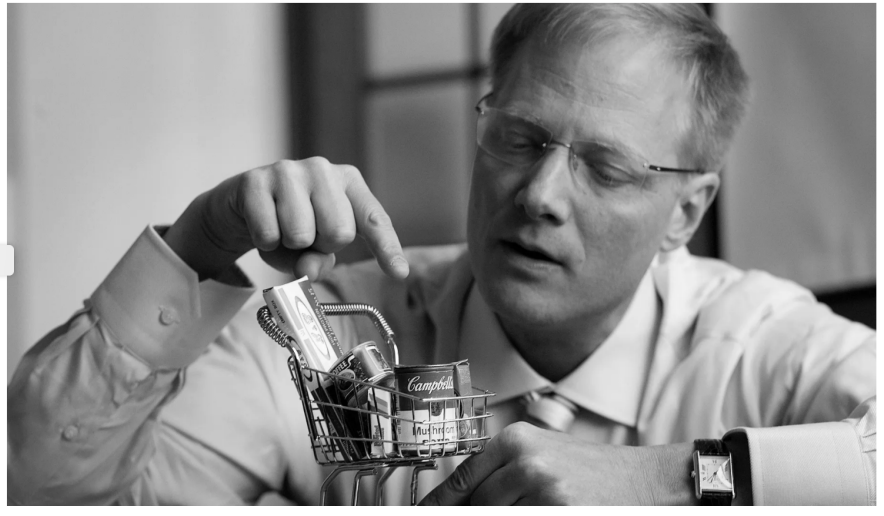
Archived post here.

A top Cornell food researcher has had 15 studies retracted. That's a lot.

Brian Wansink is a cautionary tale in bad incentives in science.

By Brian Resnick and Julia Belluz | Updated Oct 24, 2018, 2:25pm EDT

SHARE

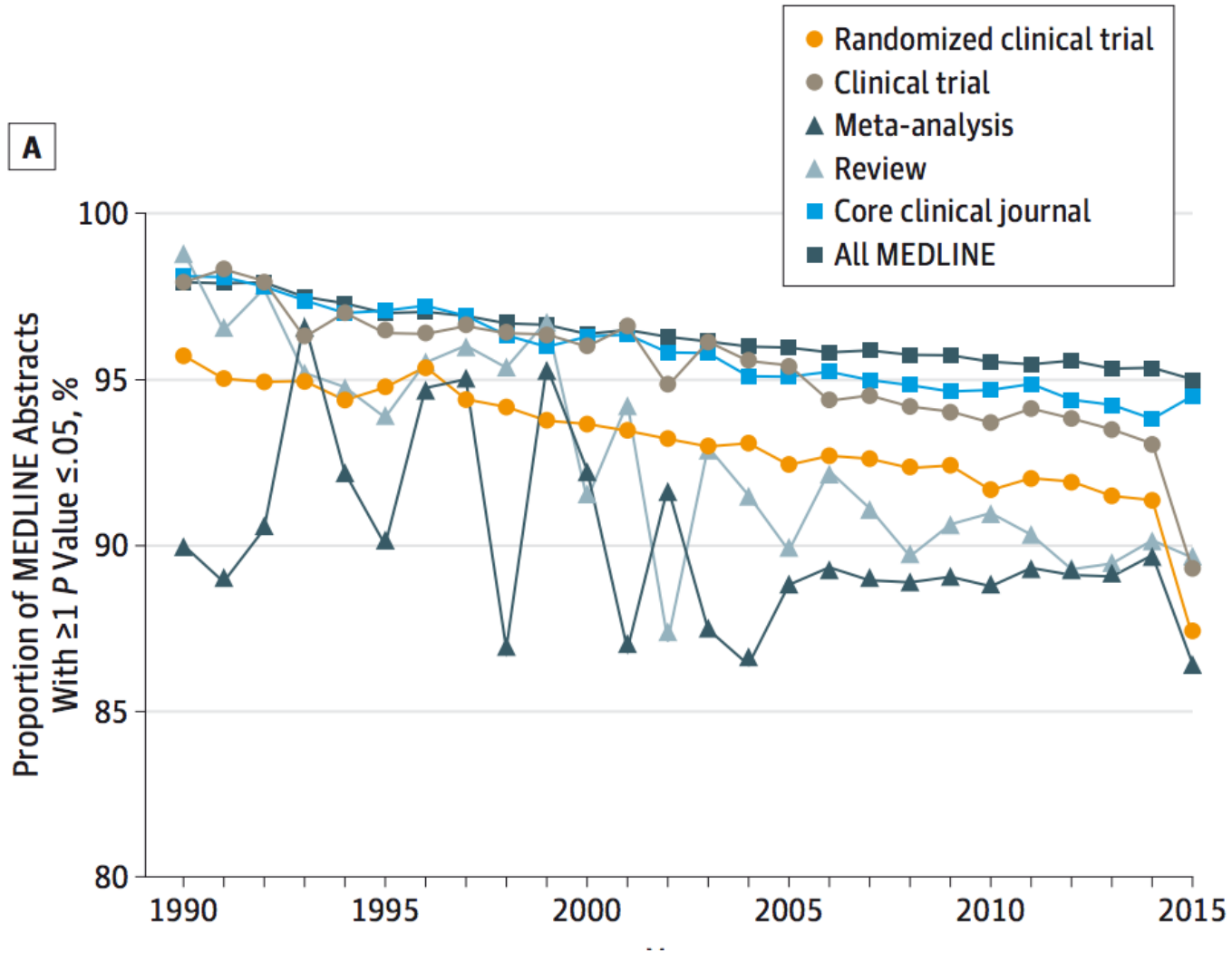


Wansink resigned from Cornell in 2019.

How do we know there is p-hacking?

(2) Seriously, everything is significant

P-values in the biomedical literature, 1990-2015



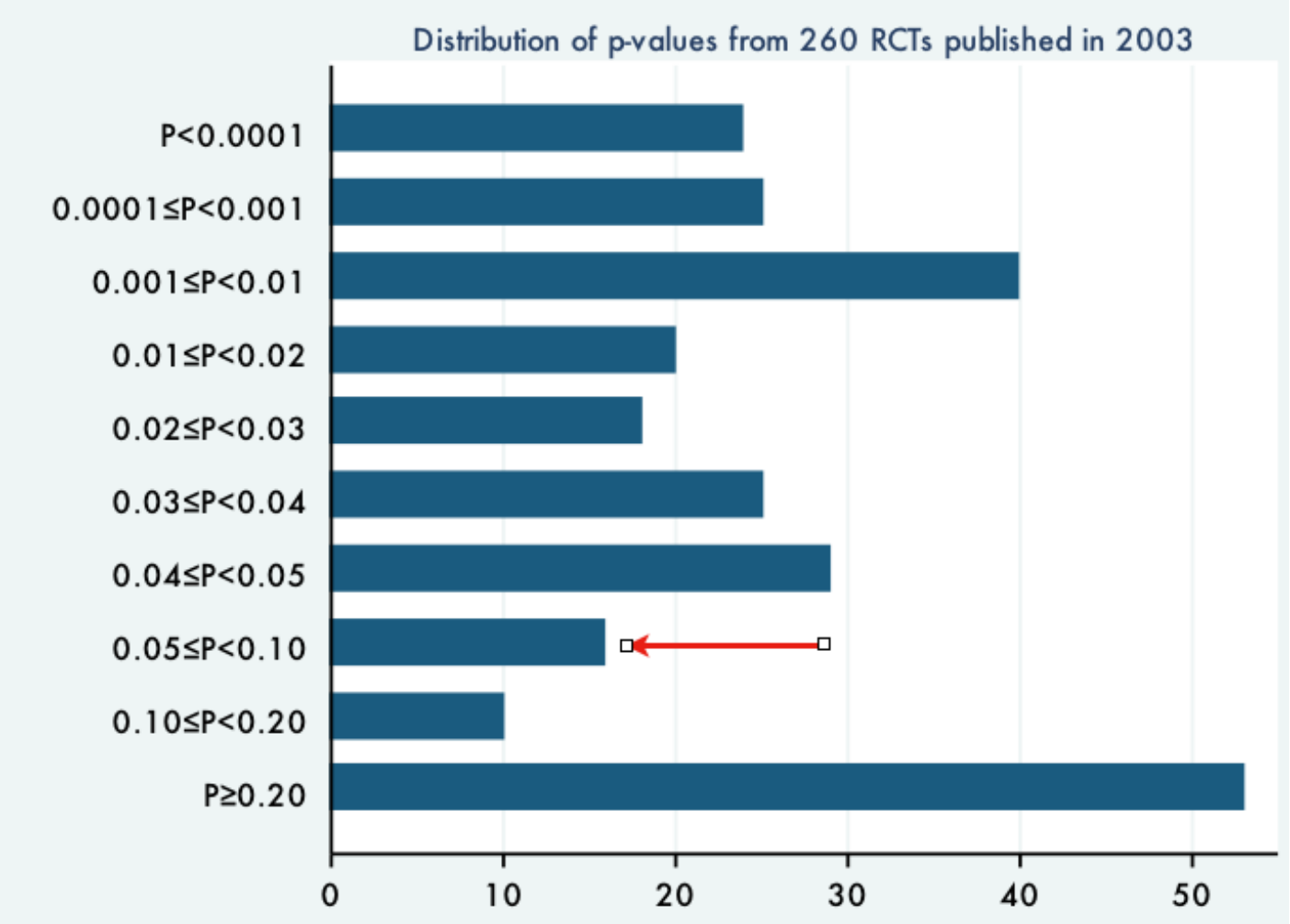
Chavalarais et al. (2013)

How do we know there is p-hacking?

(3) Maldistribution of published p-values

True for medicine, economics, psychology, political science, many other disciplines.

P-values from 260 RCTs

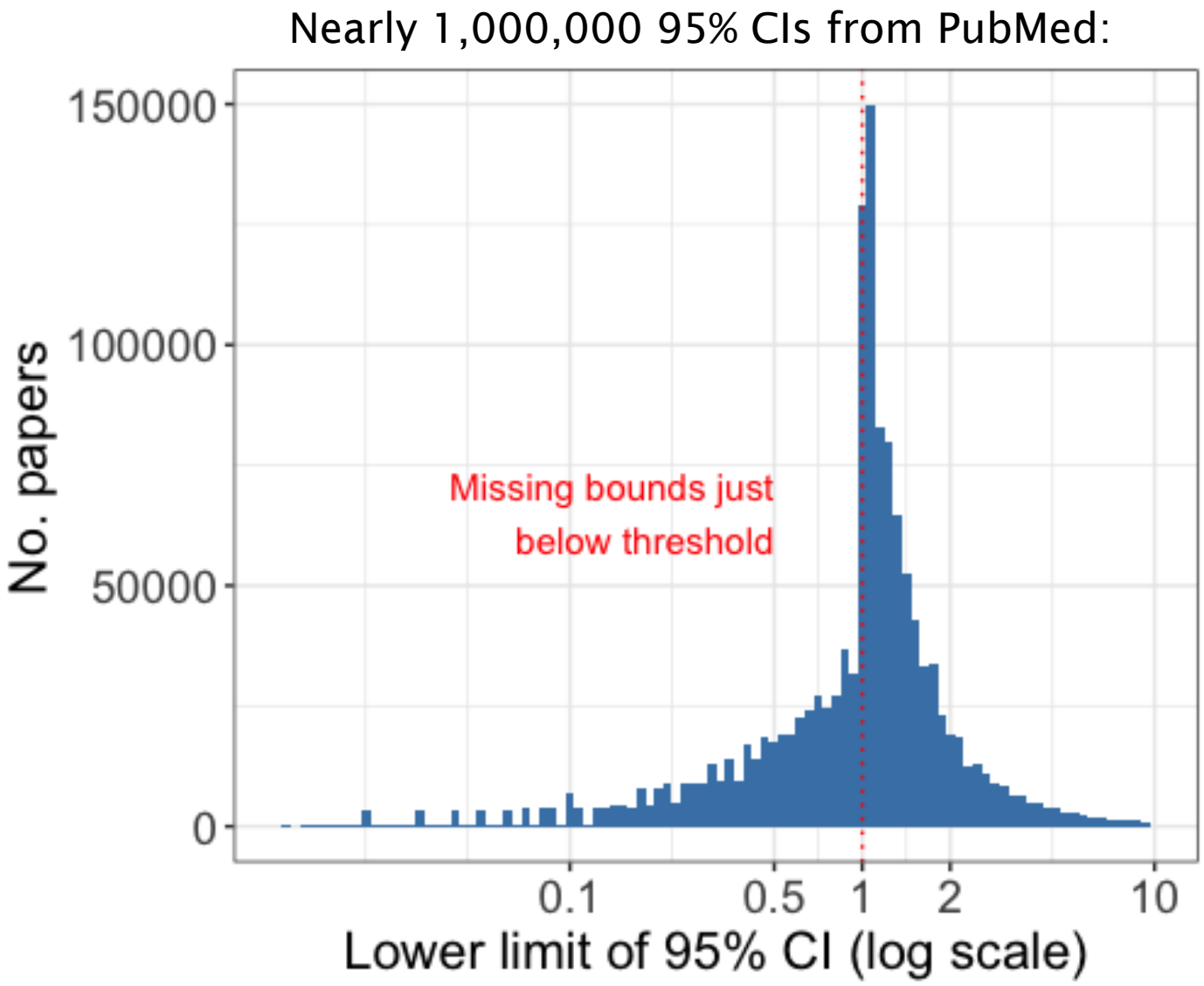


Gotzsche (2006)

Won't 95% confidence intervals help?

No.

Researchers still dichotomize them.



data from Barnett and Wren (2019)

NHST also leads to missing evidence and publication bias

Missing evidence

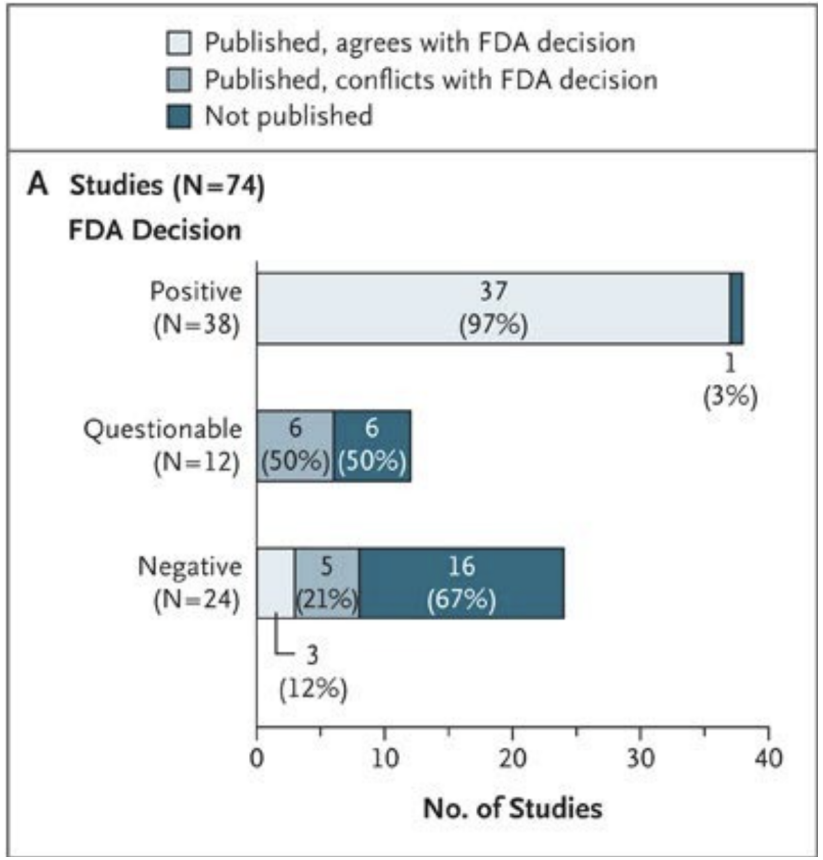
Negative studies of antidepressents less likely to be published.

Impacts regulatory decisions.

SPECIAL ARTICLE

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.



Turner et al. *NEJM* (2008)

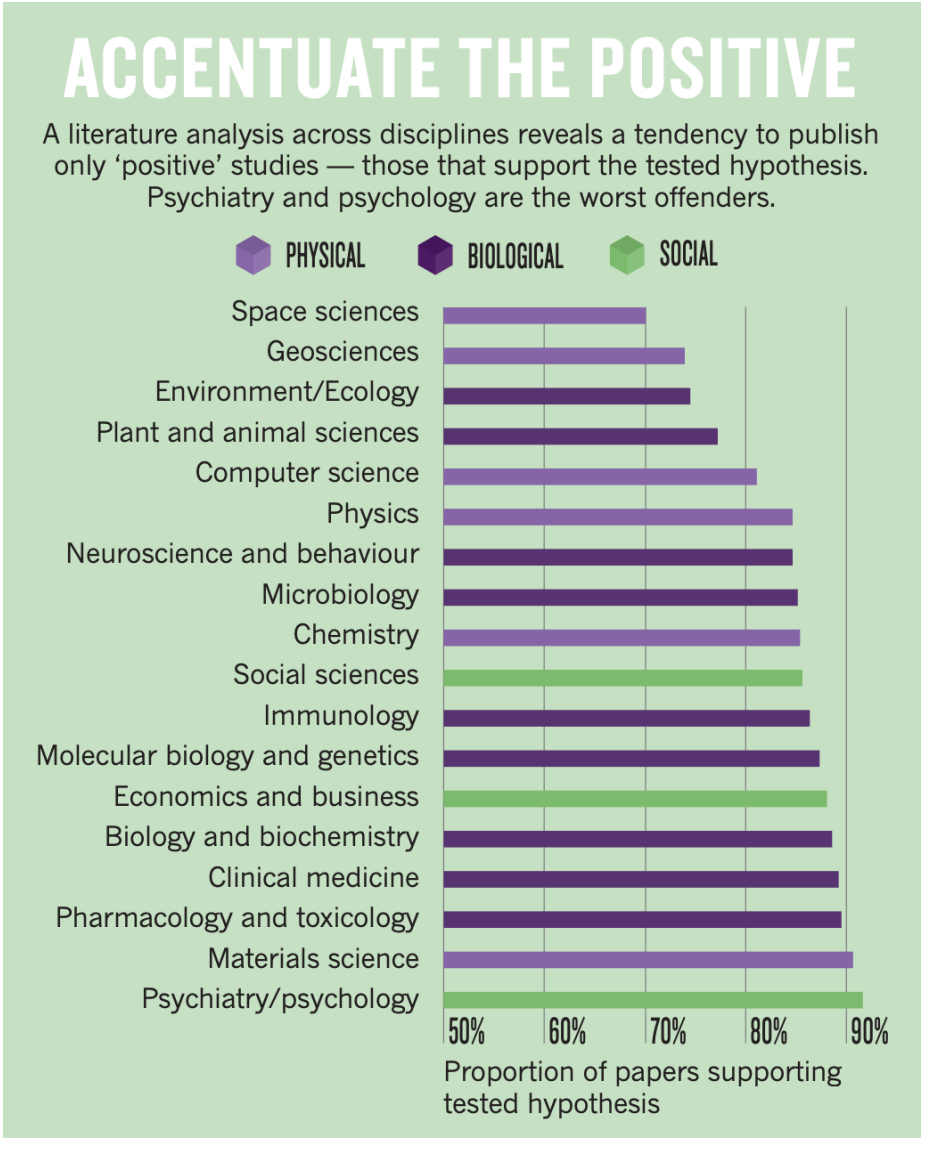
Publication bias affects nearly all disciplines

Statistically significant results are more likely to be published, across virtually all disciplines.

May be worse in "softer" sciences.

Much of the bias is likely self-imposed.

Fanelli *PLoS ONE* (2010), Yong *Nature* (2012)



Self-imposed by many researchers

221 survey experiments funded by US NSF.

All peer reviewed, required to be deposited in a registry.

All studies had results.

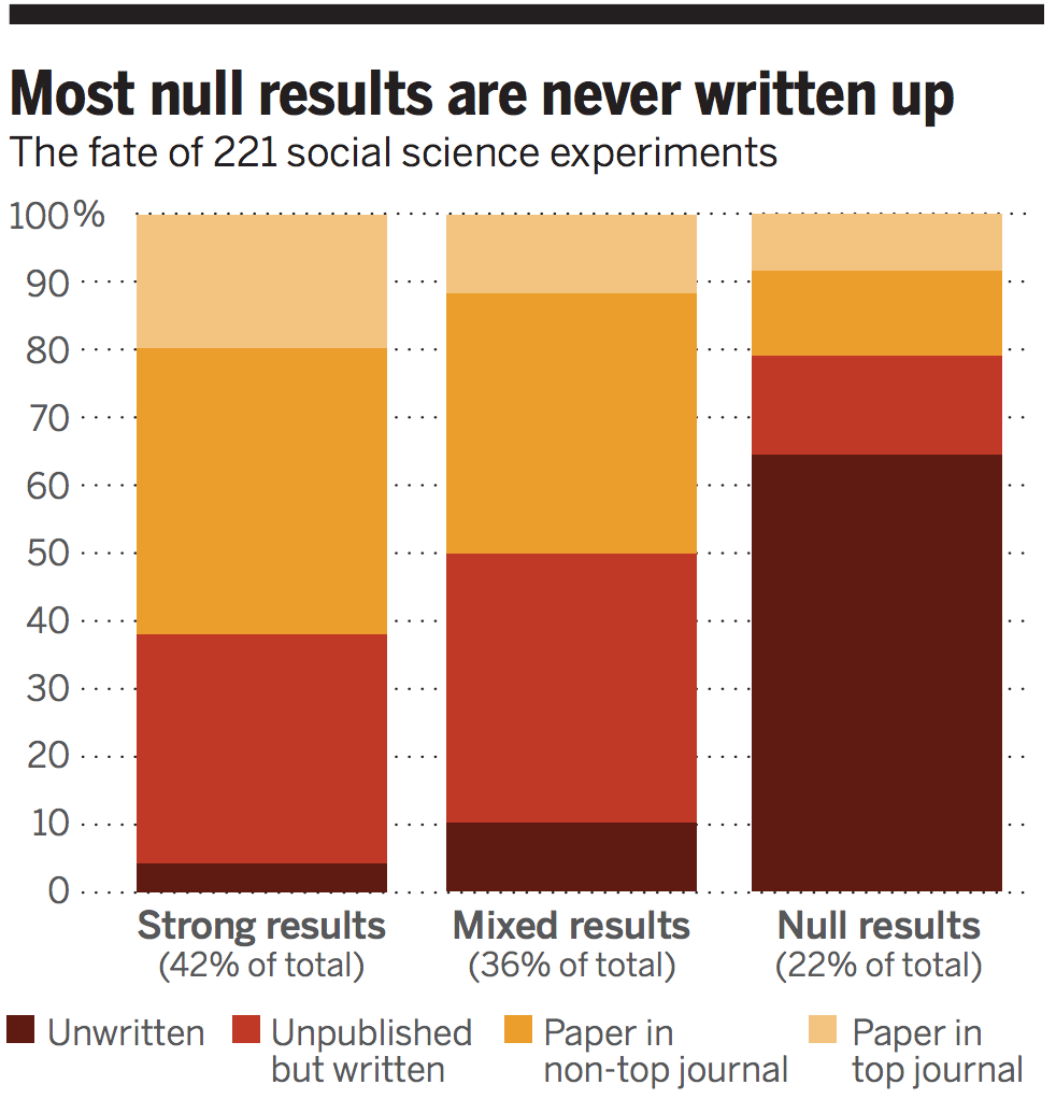


Figure from Mervis in Science 29 Aug 2014;345:992

What about peer review?

Peer review is:

- Slow, inefficient, and expensive.
- Reviewers agreement no better than chance.
- Does not detect errors.

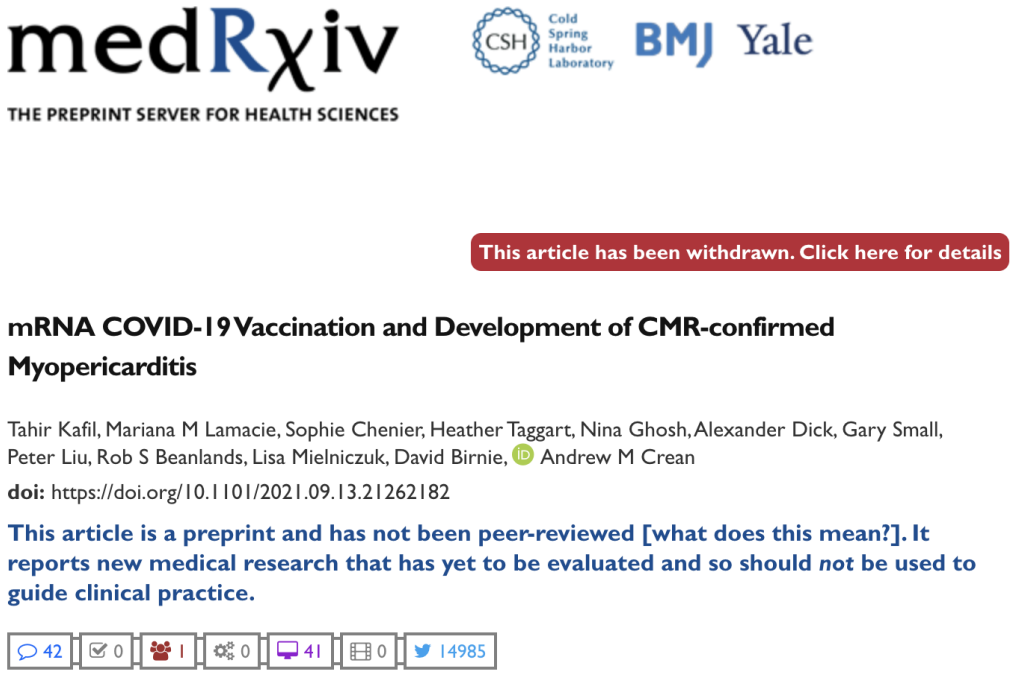
Reviewiers are biased against:

- Less prestigious institutions.
- Against new or original ideas.

Smith (2010), editor at *BMJ* for many years.

Okay, forget peer review

- Study *preprints* get science out rapidly, before peer review.
- Explosion of preprints during COVID-19
- 12 authors from Ottawa Heart Institute published preprint claiming 1 in 1000 vaccine recipients had myocarditis.
  - Case series
  - Wrong denominators (33,000 instead of 850,000)
  - Revised risk 1 in 25,000



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

This article has been withdrawn. Click here for details

mRNA COVID-19 Vaccination and Development of CMR-confirmed Myopericarditis

Tahir Kafil, Mariana M Lamacie, Sophie Chenier, Heather Taggart, Nina Ghosh, Alexander Dick, Gary Small, Peter Liu, Rob S Beanlands, Lisa Mielniczuk, David Birnie, Andrew M Crean

doi: https://doi.org/10.1101/2021.09.13.21262182

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

42 | 0 | 1 | 0 | 41 | 0 | 14985

See https://www.cbc.ca/news/health/covid-19-vaccine-study-error-anti-vaxxers-1.6188806

If we wanted to reproduce, often the materials aren't there

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa

Abstract

A reproducibility crisis is a situation where many scientific studies cannot be reproduced. Inappropriate practices of science, such as HARKing, p-hacking, and selective reporting of positive results, have been suggested as causes of irreproducibility. In this editorial, I propose that a lack of raw data or data fabrication is another possible cause of irreproducibility.

As an Editor-in-Chief of *Molecular Brain*, I have handled 180 manuscripts since early 2017 and have made 41 editorial decisions categorized as "Revise before review," requesting that the authors provide raw data. Surprisingly, among those 41 manuscripts, 21 were withdrawn without providing raw data, indicating that requiring raw data drove away more than half of the manuscripts. I rejected 19 out of the remaining 20 manuscripts because of insufficient raw data. Thus, more than 97% of the 41 manuscripts did not present the raw data supporting their results when requested by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases.

Considering that any scientific study should be based on raw data, and that data storage space should no longer be a challenge, journals, in principle, should try to have their authors publicize raw data in a public database or journal site upon the publication of the paper to increase reproducibility of the published results and to increase public trust in science.

Keywords: Raw data, Data fabrication, Open data, Open science, Misconduct, Reproducibility

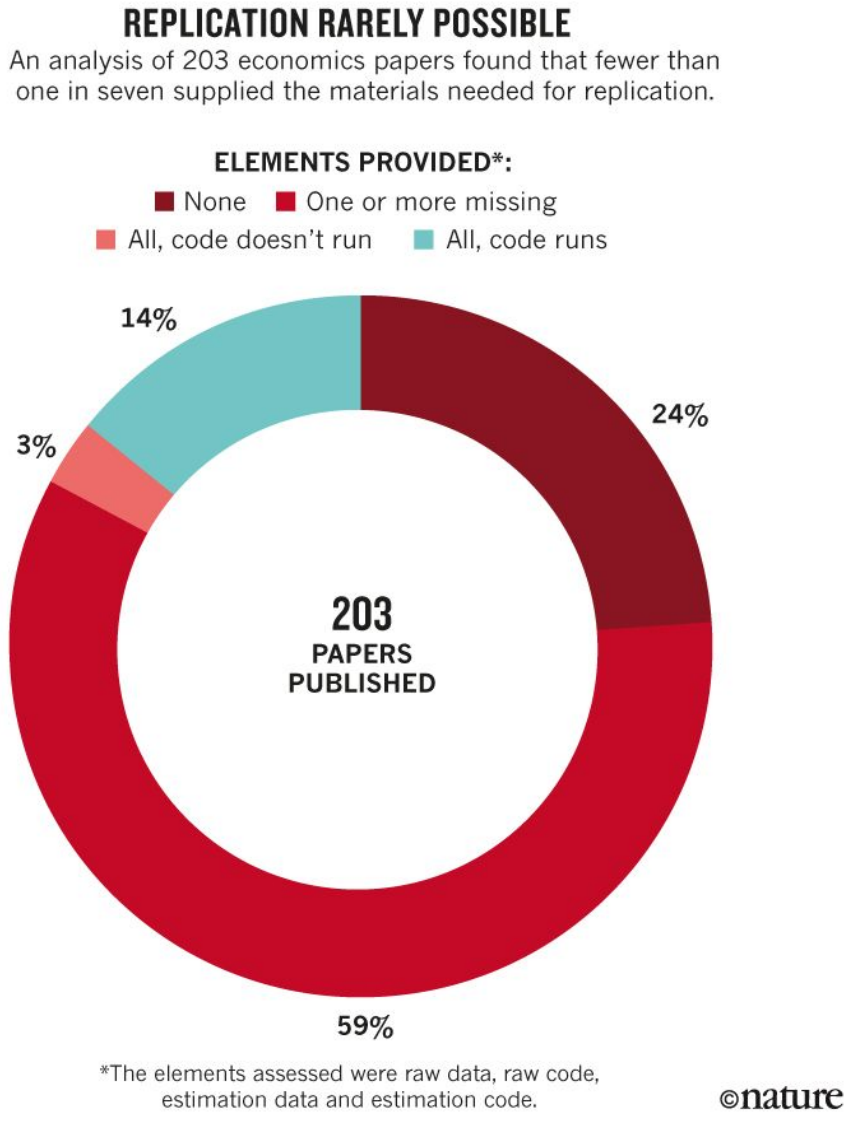
Miyakawa *Molecular Brain* (2020) 13:24

Even with data, efforts to reproduce are rarely successful

Gertler et al. gathered replication materials from published papers in econ.

Most authors only included estimation code.

Estimation code only ran in 40% of cases.



Gertler et al. 2018

Outline

What is the replication crisis?

What do we mean by replication?

What are some potential solutions?

Preregistration of studies

What is study preregistration?

A detailed study proposal that is:

Time stamped

Records and publicizes time and date.

Read-only

Can't be modified.

Registered prior to data collection/access

Robust to fieldwork, data snooping.

What is preregistration?

Common / required for publishing most RCTs

Controversial for observational studies.

Idea is to help *reduce publication bias*, since registered studies may be followed over time.

No guarantee anyone will publish.

Also can provide intellectual provenance of your ideas and hypotheses.

Good for planning and hypothesizing, not a straightjacket.

Why preregistration?

1. It's *not* about minimizing Type 1 errors.

2. It *is* about:

Allowing others to transparently evaluate the credibility of the analysis.

Assuring that all of the evidence is available for synthesis.

Why does preregistration matter?

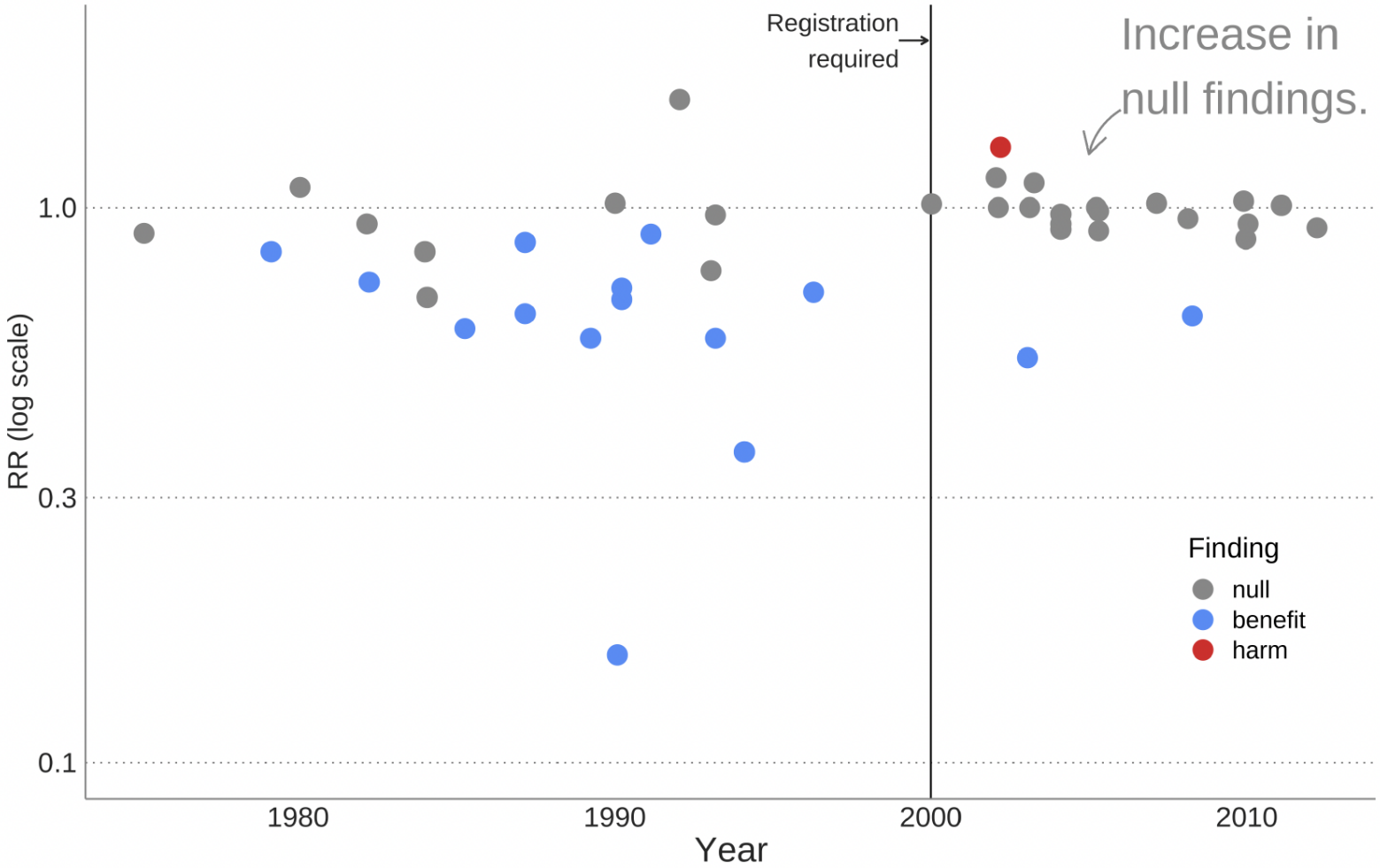
Evidence synthesis should be on *all* the evidence.

Distorts planning of future studies.

Unethical and wasteful.

Registration is useful

In 2000 NHLBI required the registration of primary outcome on ClinicalTrials.gov for all their grant-funded activity.

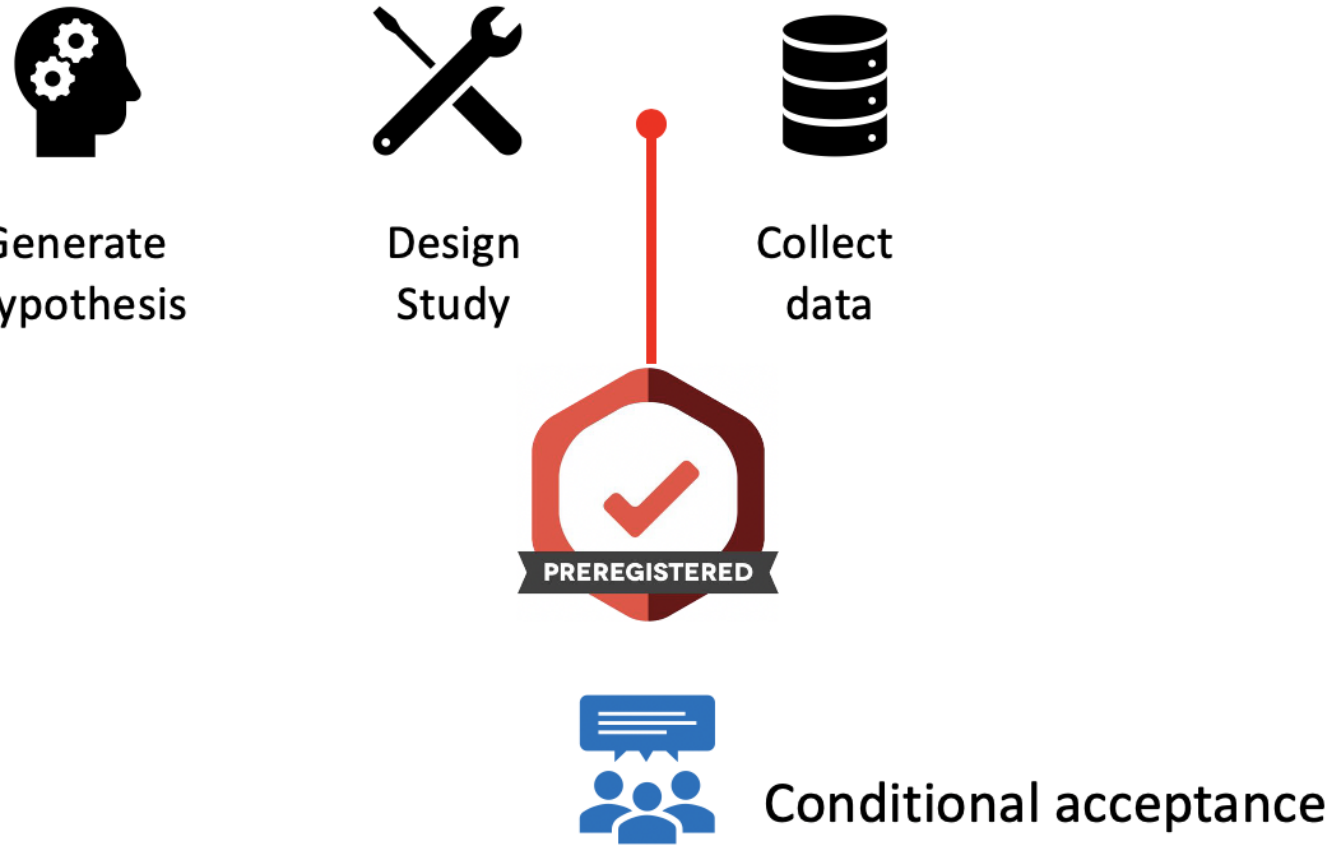


redrawn from Kaplan and Irwin (2015)

What if my results are null?

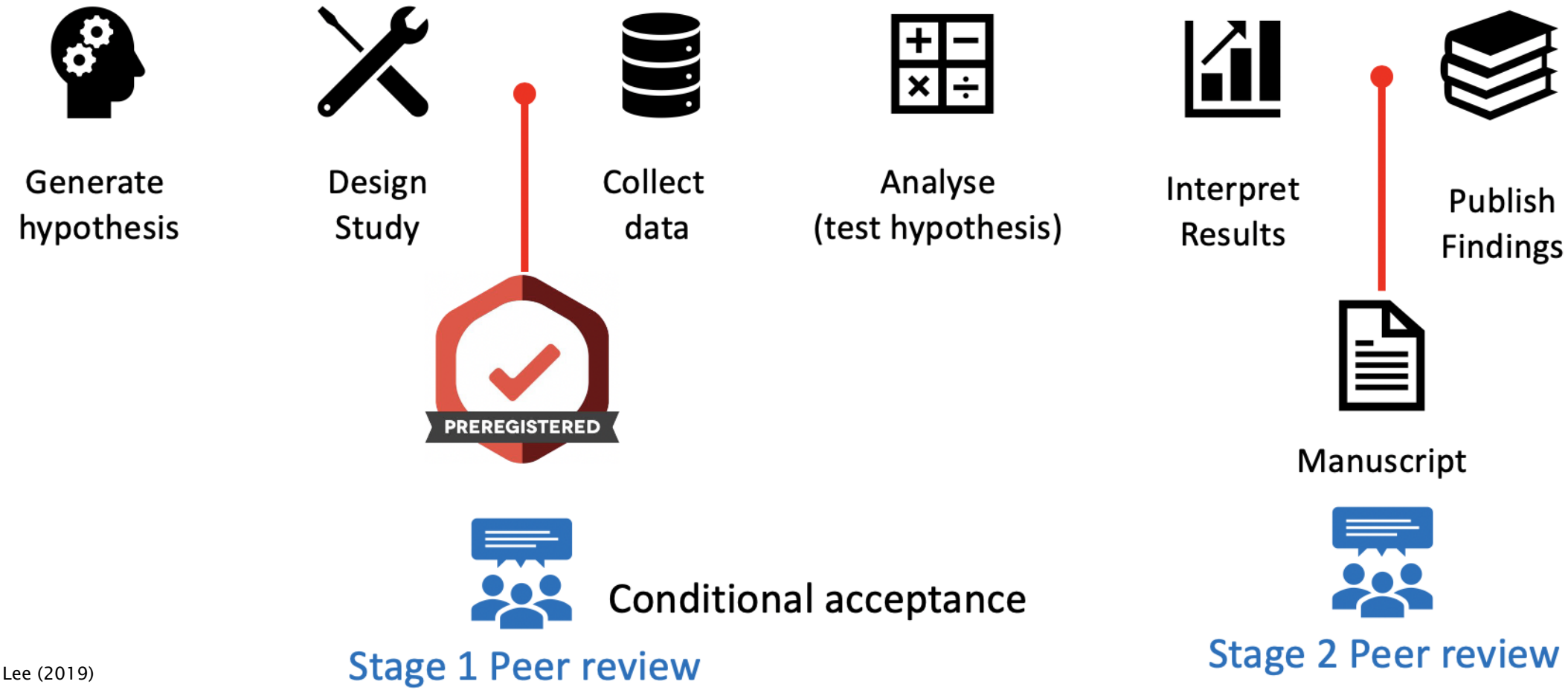
You showed us that they won't get published!

Emphasis on design: Registered Reports



Lee (2019)

Emphasis on design: Registered Reports

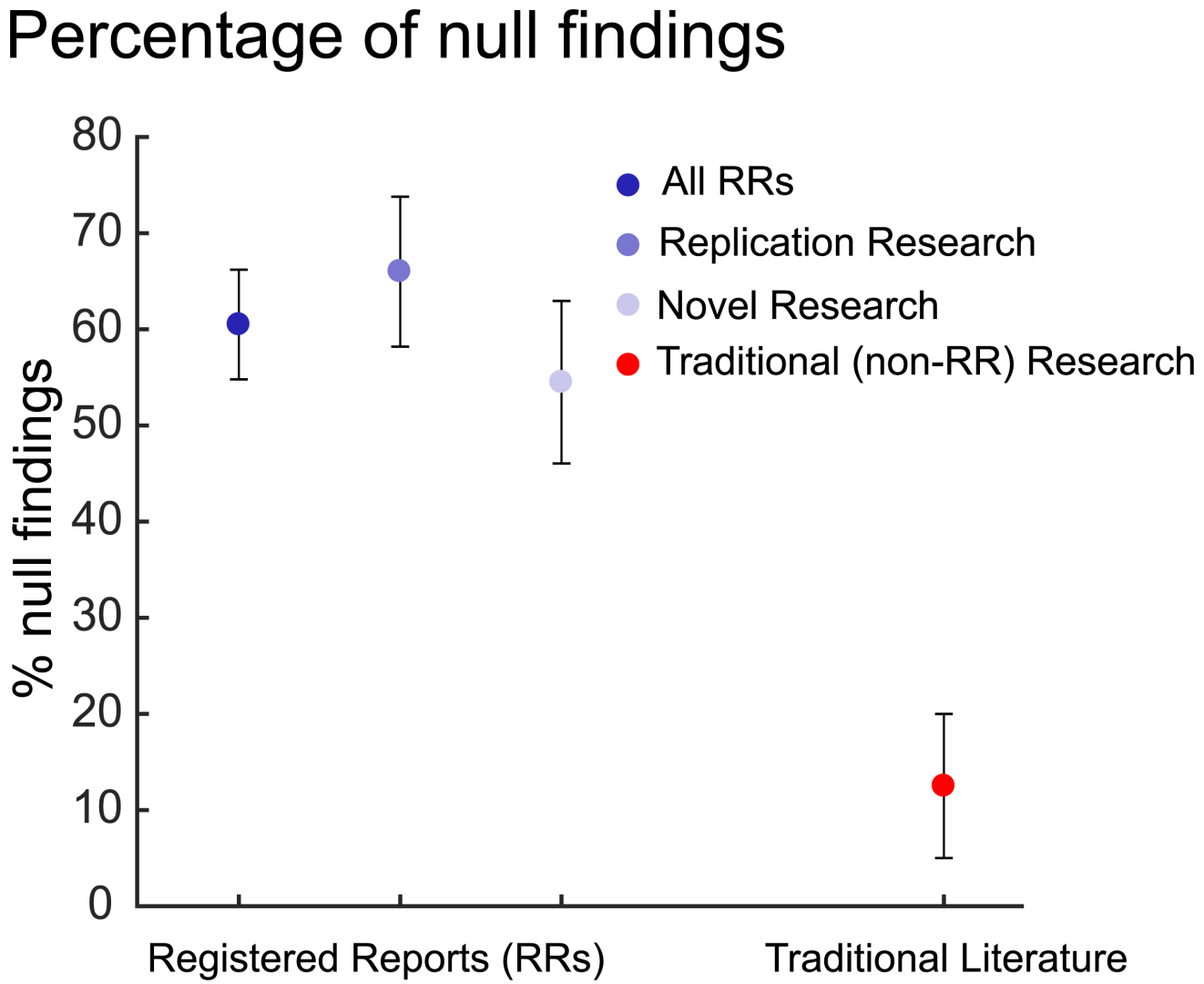


Lee (2019)

RRs in Psychology

Little difference between 'replication' studies and 'novel' studies.

Big difference from non-registered studies.

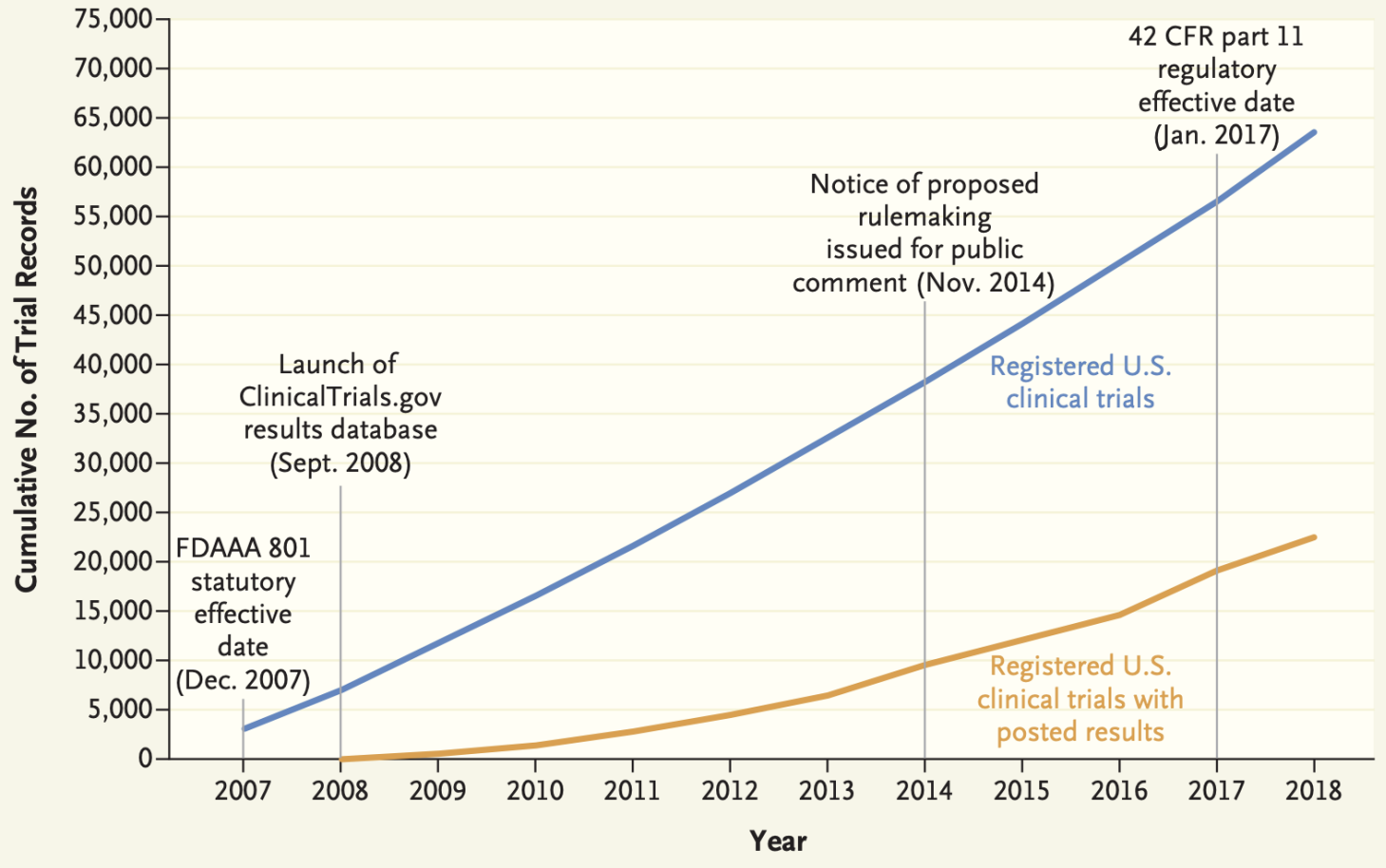


Allen & Mehler, *PLoS Biology* (2019)

Registration is useful

but not sufficient

A majority of (pre-COVID-19) registered RCTs still not reported.



Average No. of Records/Wk	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Trials completed	58	79	89	92	99	103	104	111	112	122	120	135
Results posted	0	2	10	16	24	33	40	61	46	50	86	68

Zarin *NEJM* (2019)

But is preregistration enough?

- Still many differences between registration and published reports.

RESEARCH Open Access

COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time

Ben Goldacre[1*], Henry Drysdale[1], Aaron Dale[1], Ioan Milosevic[1], Eirion Slade[1], Philip Hartley[1], Cicely Marston[2], Anna Powell-Smith[1], Carl Heneghan[1] and Kamal R. Mahtani[1]

Methods

We set out to prospectively identify all trials published in five leading medical journals over a six-week period, identify every correctly and incorrectly reported outcome in every trial by comparing the published report against the published pre-trial protocol (or, where this was unavailable, the pre-trial registry entry), write a correction letter to the journal for publication on all misreported trials, and document the responses from journals. We used mixed methods combining quantita-

Goldacre (2019)

Academic journals are not helping

Summary statistics on correction letter publication

	Annals	BMJ	JAMA	*Lancet*	NEJM	Total
Letters required	5	2	11	20	20	58
Percentage of letters required	100.00%	66.70%	84.60%	83.30%	90.90%	86.6% (95% CI 78.4–94.7%)
Letters published	5	2	0	16	0	23
Percentage of letters published	100%	100%	0%	80%	0%	39.7% (95% CI 27.0%–53.4%)
Mean publication delay for published letters	0 days (online)	0 days (online)	n/a	150 days	n/a	104 days (median 99 days, range 0–257 days)

Abbreviations: *BMJ British Medical Journal*, *CI* confidence interval, *CONSORT* Consolidated Standards of Reporting Trials, *JAMA Journal of the American Medical Association*, *n/a* not applicable, *NEJM New England Journal of Medicine*

Goldacre (2019)

Preregistration is not a panacea

Preregistered $\neq$ correct/sensible/useful

Transparency helps, but cannot fix terrible design or methods.

Post-hoc analysis can be worthwhile

Probing surprising results or mechanisms generates knowledge.

May also lead to 'halo' effects

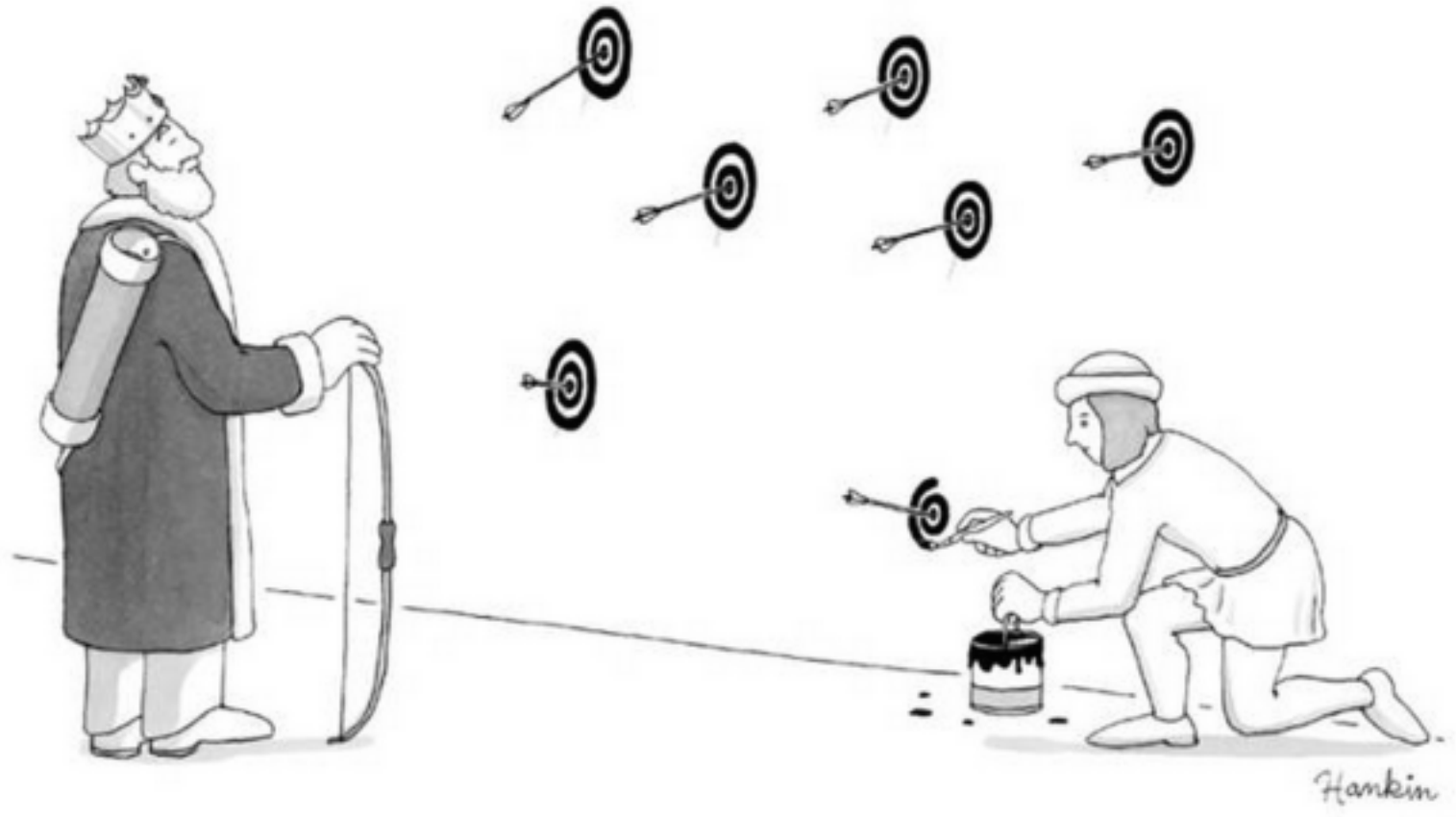
Preregistered research deserves equal opportunity interrogation.

See Nosek et al. (2019) and Szollosi et al. (2020)

Pre-analysis plans

Hypothesizing **A**fter the **R**esults are **K**nown (HARKing)

- Pretending what you found was what you were looking for.
- Easy to "find" theory / biological evidence consistent with results.



New Yorker, 2014-12-07

What is a pre-analysis plan?

- Detailed description of research design and data analysis plans, submitted to a registry before looking at the data.
- Helps to tie your hands for data analysis (address researcher degrees of freedom, etc.).
- Distinguish between confirmatory and exploratory analysis.
- Increases the credibility of research.
- Transparent methods make it easier for others to build on your work.

Confirmatory and exploratory studies have different aims

Confirmatory

- Well-theorized.
- Plausible mechanisms.
- Minimize false positives.
- Hypothesis *testing*.

Exploratory

- Pushes new ideas.
- Hypothesis *generating*
- Minimize false negatives.
- Testing irrelevant.

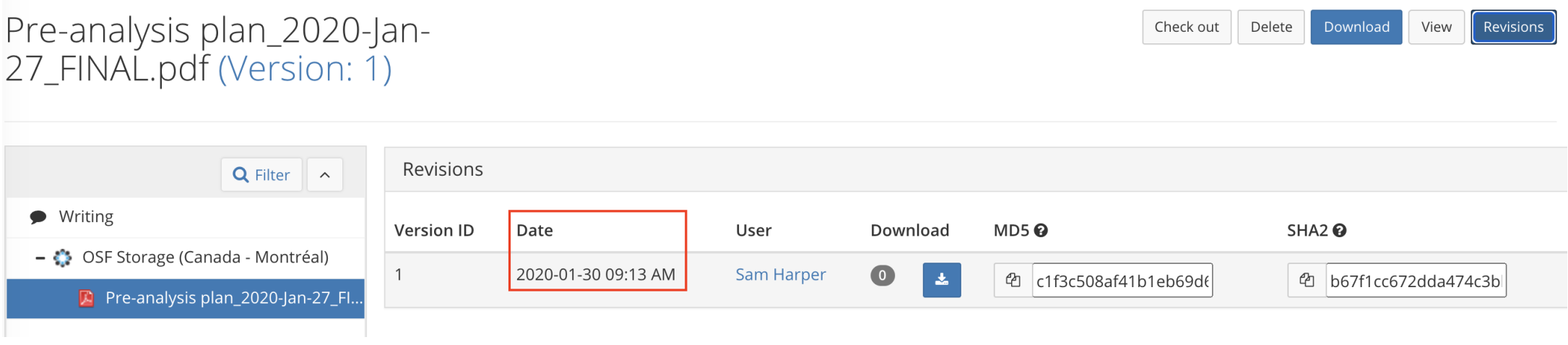
Dehaven and Bowman OSF (2020)

What goes into a pre-analysis plan?

- General info (Title, PIs, Staff)
- Introduction and Summary
- Study Design:
    - Hypotheses
    - Main variables
    - Study setting.
    - Intervention components.
    - Data collection methods.
    - Treatment assignment mechanism.
    - Power calculations.
- Analytic decisions
    - models
    - derived variables
    - clustering
    - multiple testing
- Threats/mitigation/robustness checks.
- Dissemination plans

Example from epidemiology

Note the time-stamp, which provides credible evidence of *when* you had your brilliant ideas.



Example from epidemiology

Can be challenging for observational studies or secondary data analyses.

Can you prove when you obtained data access?

Pre-analysis plan for "Short term benefits but long term harm? Assessing the consequences of antenatal corticosteroid administration for child neurodevelopment"

Jennifer A Hutcheon[1], Sam Harper[2], Amanda Skoll[1], Myriam Srour[3], Jessica Liauw[1], Erin Strumpf[2,4]

[1]Department of Obstetrics & Gynaecology, University of British Columbia
[2] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University
[3] Department of Pediatric Neurology, McGill University
[4] Department of Economics, McGill University

Purpose

This document describes a pre-analysis plan for a study examining the child health consequences of antenatal corticosteroid administration in a population-based cohort of linked administrative and clinical records from British Columbia, Canada. We use a regression discontinuity design that exploits the pronounced change in antenatal corticosteroid administration practices based on a clinical practice guideline that recommended administration up to 33 weeks, 6 days of gestation (33+6 weeks), but not at or beyond 34+0 weeks. This pre-analysis plan was written after the individual datasets had been received and some descriptive statistics calculated for key variables, but prior to the linkage of the datasets or analyses linking exposure with longer-term child health outcomes.

2. Design Solutions

2.1 Preregistration

2.2 Pre-analysis plans

2.3 Reporting guidelines

"Most publications have elements that are missing, poorly reported, or ambiguous"

Glasziou et al. (2014)

Reducing waste from incomplete or unusable reports of biomedical research

Paul Glasziou, Douglas G Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, Elizabeth Wager

Abstract **Trials:** missing effect size and confidence interval (38%); no mention of adverse effects (49%)[72]
Methods **Trials:** 40–89% inadequate treatment descriptions[11, 13] **fMRI studies:** 33% missing number of trials and durations[3] **Survey questions:** 65% missing survey or core questions[25] **Figures:** 31% graphs ambiguous[45]
Results **Clinical trials:** outcomes missing: 50% efficacy and 65% harm outcomes per trial incompletely reported[6] **Animal studies:** number of animals and raw data missing[17] (54%, 92%); age and weight missing (24%) **Diagnostic studies:** missing age and sex (40%)[15]
Discussion **Trials:** no systematic attempt to set new results in context of previous trials (50%)[69]
Data **Trials:** most data never made available; author-held data lost at about 7% per year

Figure **3: Estimates of the prevalence of some reporting problems (see publication column, figure 1).**
fMRI=functional MRI.

Importance of intervention details

Want decision-makers to act on your evidence?

Can they actually understand what you did?

Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials


OPEN ACCESS

Tammy C Hoffmann *associate professor of clinical epidemiology*, Chrissy Erueti *assistant professor*, Paul P Glasziou *professor of evidence-based medicine*

- Of 137 interventions, only 53 (39%) were adequately described;
- The most frequently missing item was the “intervention materials” (47% complete);

Hoffman (2013)

Missing due to:

- copyright or intellectual property;
- absent materials or intervention details;
- unaware of their importance.

Reasons given for study intervention materials being unavailable

Category of reason (number of authors providing a response in this category) and illustrative quotes from authors:

Materials not publicly available (9)

"Due to legal copyright restrictions at my university I am unable to send"

"Not publicly available because we based them on materials provided by our local government"

"Not publicly available—only to our trainers"

"Not yet—they will be made publicly available within two years"

"No it is not. Attached is a table of contents"

"The training materials from the trial are not online—we had no real reason to do that"

Corresponding author did not have copy of materials to send or could not provide further details about intervention (8)

"People originally in the position have moved on"

"I am unable to find . . . my old computer files"

"I'm afraid I no longer have access to those materials"

"I do not have it"

"I am not able to answer most of your questions. I was not involved with running the trial, only analysing and reporting on the QOL results after the data was collected"

"I can't provide these"

Other (3)

"You will have to read the literature"

"No, is in Dutch"

"The [materials] are tailored, thus it is difficult to disseminate. We could send an example"

Materials were previously publicly available but no longer are (2)

"URL doesn't exist anymore"

"We had been making it previously available, but need to update it, so are no longer"

Hoffman (2014)

Reporting guidelines exist for entire research lifecycle

Question and approach	Systematic review 👉 PRISMA/PROSPERO
Pre-intervention	Research protocol/preanalysis 👉 SPIRIT
Research report	Trials/Observational studies 👉 CONSORT/STROBE
Cost-effectivness	Benefits and costs of interventions 👉 CHEERS

See the EQUATOR Network

equator network

Enhancing the QUAlity and Transparency Of health Research

EQUATOR resources in German | Portuguese | Spanish

Home | About us | Library | Toolkits | Courses & events | News | Blog | Librarian Network | Contact

Your one-stop-shop for writing and publishing high-impact health research

find reporting guidelines | improve your writing | join our courses | run your own training course | enhance your peer review | implement guidelines

Library for health research reporting

The Library contains a comprehensive searchable database of reporting guidelines and also links to other resources relevant to research reporting.

- Search for reporting guidelines
- Not sure which reporting guideline to use?
- Reporting guidelines under development
- Visit the library for more resources

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	
Economic evaluations	CHEERS	

See all 442 reporting guidelines

(Some) evidence that it might matter.

- Some evidence that item reporting has increased.
- Consistent with revised CONSORT (2001).
- Non-adopting journals report fewer items.

The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed

Sally Hopewell, senior research fellow,[1] Susan Dutton, senior medical statistician,[1] Ly-Mee Yu, senior medical statistician,[1] An-Wen Chan, assistant professor,[2] Douglas G Altman, director [1]

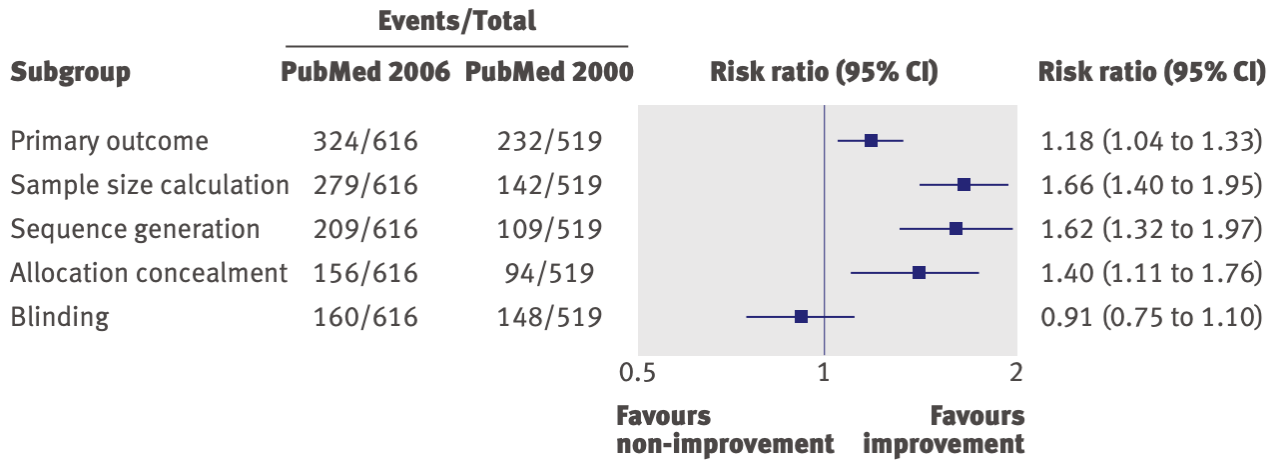


Fig 2 | **Differences in reporting of methodological items between 2000 and 2006**

Hopewell et al. *BMJ* (2010)

Since 2015 funders, journals are embracing *Transparency and Openness* (TOP) guidelines.

8 MODULAR STANDARDS

Citation Standards Describes citation of data	**Data Transparency** Describes availability and sharing of data
Analytical Methods Transparency Describes analytical code accessibility	**Research Materials Transparency** Describes research materials accessibility
Design and Analysis Transparency Sets standards for research design disclosures	**Preregistration of Studies** Specification of study details before data collection
Preregistration of Analysis Plans Specification of analytical details before data collection	**Replication** Encourages publication of replication studies

ACROSS 3 TIERS

1 **DISCLOSURE:** the final research output must disclose if the work satisfies the standard

2 **REQUIREMENT:** the final research output must satisfy the standard

3 **VERIFICATION:** third party must verify that the standard is being met

Source: OSF

It's still difficult to change norms

Most journals chose *Level 1* (disclosure)

J Am Heart Assoc published 40 original research papers during first half of 2019.

- Posted data: 0
- Posted code: 1
- Data upon "reasonable" request: 30
- Code upon "reasonable" request: 5

MINI-REVIEW

American Heart Association.

Resource Sharing to Improve Research Quality

Ghassan B. Hamra, PhD; Neal D. Goldstein, PhD; Sam Harper, PhD

Transparency and openness are vital for strengthening the scientific process. However, there is no clear agreement in the scientific community about the elements necessary to qualify scientific research as a transparent and open process. Historically, the description of study methods and results within individual academic publications has been treated as sufficient for establishing transparency; that is, based solely on the written description of study procedures and analytic techniques, a third party can be *assumed* to have all the information needed to reproduce the results of an individual study if the data were available. The core philosophy of *reproducible* research is slightly different and challenges this assumption. Rather than relying on the written report, reproducible research culture demands access to data and analytic code used to produce study results. In this scenario, anyone should be able to exactly reproduce the tables, figures, and evidence presented in a given article. The push for reproducible research and current publication practices do question those findings. While self-correction is natural in science, it is not the norm,[1] and reports have suggested that the extent to which study findings cannot be replicated is alarming, leading to the so-called replication crisis.[2] Many related reasons have been put forward to explain the replication crisis, including misaligned incentives in academia, the file drawer effect,[3] p-hacking,[4] overreliance on null hypothesis significance testing,[5] and even outright falsification of data. Some have suggested that our existing assumptions about what qualifies as transparent and open in science may be insufficient and that addressing this can safeguard against further replication crises.

In this commentary, we discuss the importance of transparency and openness, focusing on the 2 major elements necessary for reproducibility: the data and analytic code used to produce the results in a published research report. We highlight how greater openness can support more reliable findings (in the long run) by allowing checks for

Hamra, Goldstein, Harper (2019)

Value of reporting guidelines

Improve transparency of reported research

Benefits funders, producers and consumers of research.

May help to improve the quality of research.

More evidence needed, unintented consequences possible.

Better reporting $\neq$ more reliable.

Tranparently reported research can still be biased/bad.

Registration, pre-analysis plans, and reporting guides are design strategies to help mitigate bias from underreported research

They do not guarantee reliable or valid research

Incentive problems

Reward structure

Papers, grants, media, "novel" and "significant" results.

Incentives

Gift authorship, CV padding, salami-slicing

Overstating claims, ignoring "non-significant" results, p-hacking

Hoarding data, non-transparent materials and methods

Summary points

Science is conducted by humans.

Extra-scientific factors matter, not necessarily malicious intent.

Changing norms is hard

Many incentives exist that undermine scientific integrity.

Transparency and openness can help

Making research process transparent is the bare minimum, and does not guarantee 'true' results.