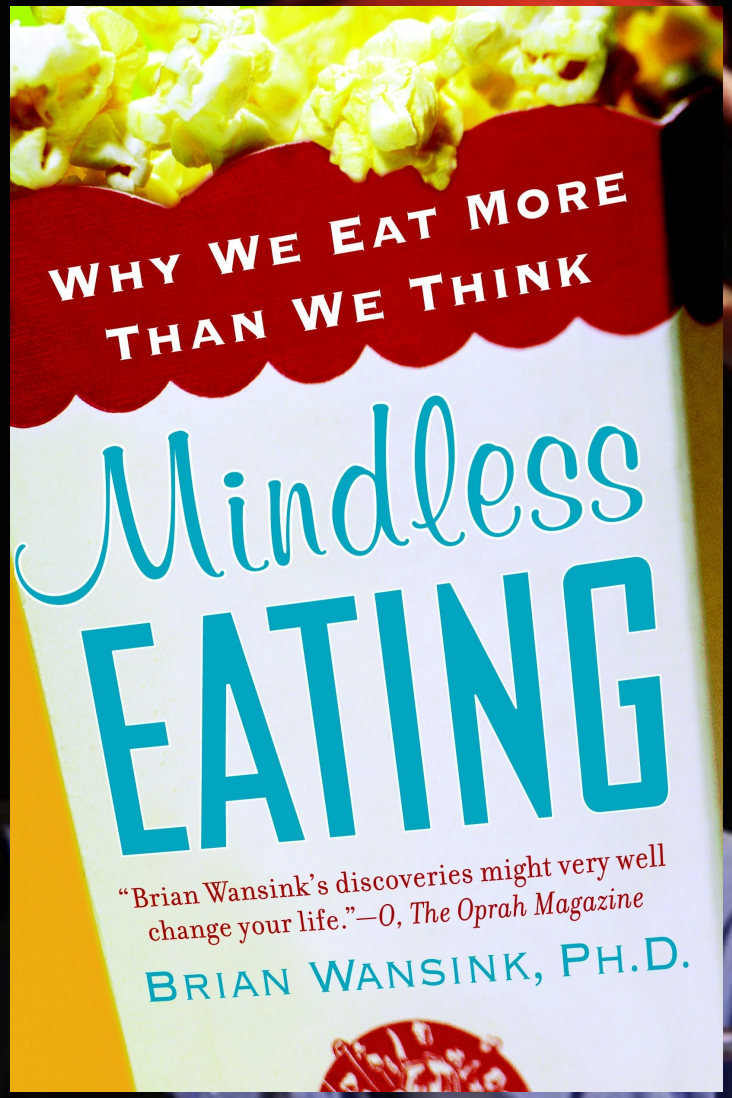
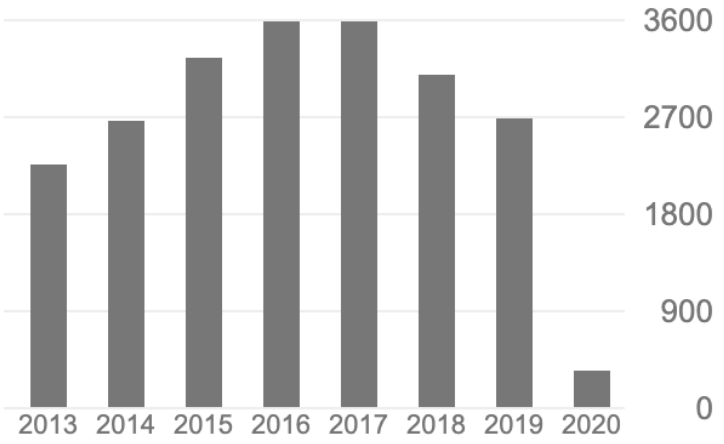
Can Open Science Improve Scientific Integrity?

Sam Harper

2020-10-29



Cited by		VIEW ALL
	All	Since 2015
Citations	31622	16656
h-index	83	64
i10-index	273	232



NOV
20
2007

Brian Wansink! At the USDA!

Every now and then something incredible happens and here it is. Brian Wansink, Cornell Professor and author of Mindless Eating, has been appointed executive director of the USDA Center for Nutrition Policy and Promotion. This is the piece of USDA responsible for dietary advice to the public. Wansink is the guy who does the terrific research on environmental determinants of overeating showing that large portions, wide drinking glasses, foods close by, and health claims encourage everyone to eat more calories than they need or want. Will he be able to anything good at USDA? Let's hope so. In the meantime, cheers to USDA for making a brilliant appointment.

https://www.foodpolitics.com/2007/11/brian-wansink-at-the-usda/

"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

Archived post here.

"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.

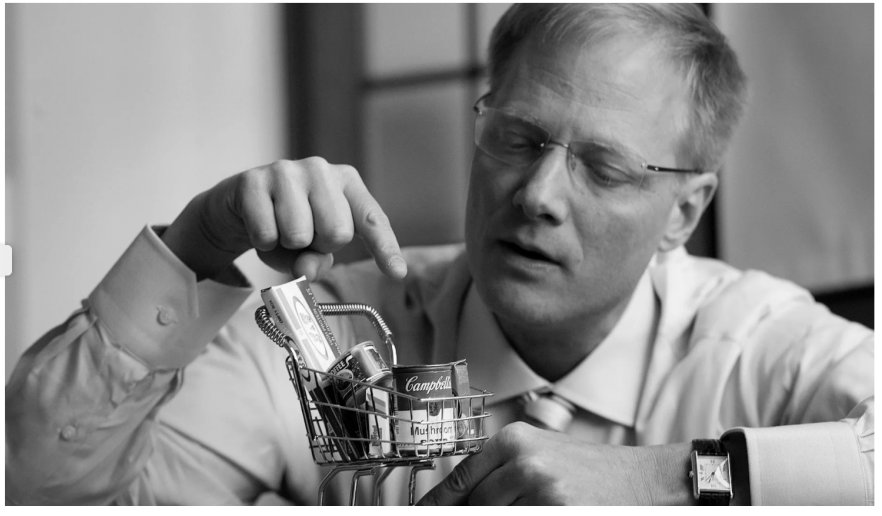
Archived post here.

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.

Why does this bother us?

Mertonian Norms and Counternorms in Science

Norms

- *Universalism*: Evaluate research only on its merit.
- *Communality*: Openly share new findings.
- *Disinterestedness*: Motivated by the desire for knowledge and discovery.
- *Skepticism*: Consider all new evidence, even if it challenges their own work.

Counternorms

- *Particularism*: New knowledge from reputation or group.
- *Secrecy*: Protect own findings for private gain.
- *Self-interestedness*: Colleagues are competitors.
- *Dogmatism*: Protecting one's own findings.

Merton (1942), Christensen et al. (2019)

How do we know that science isn't working?

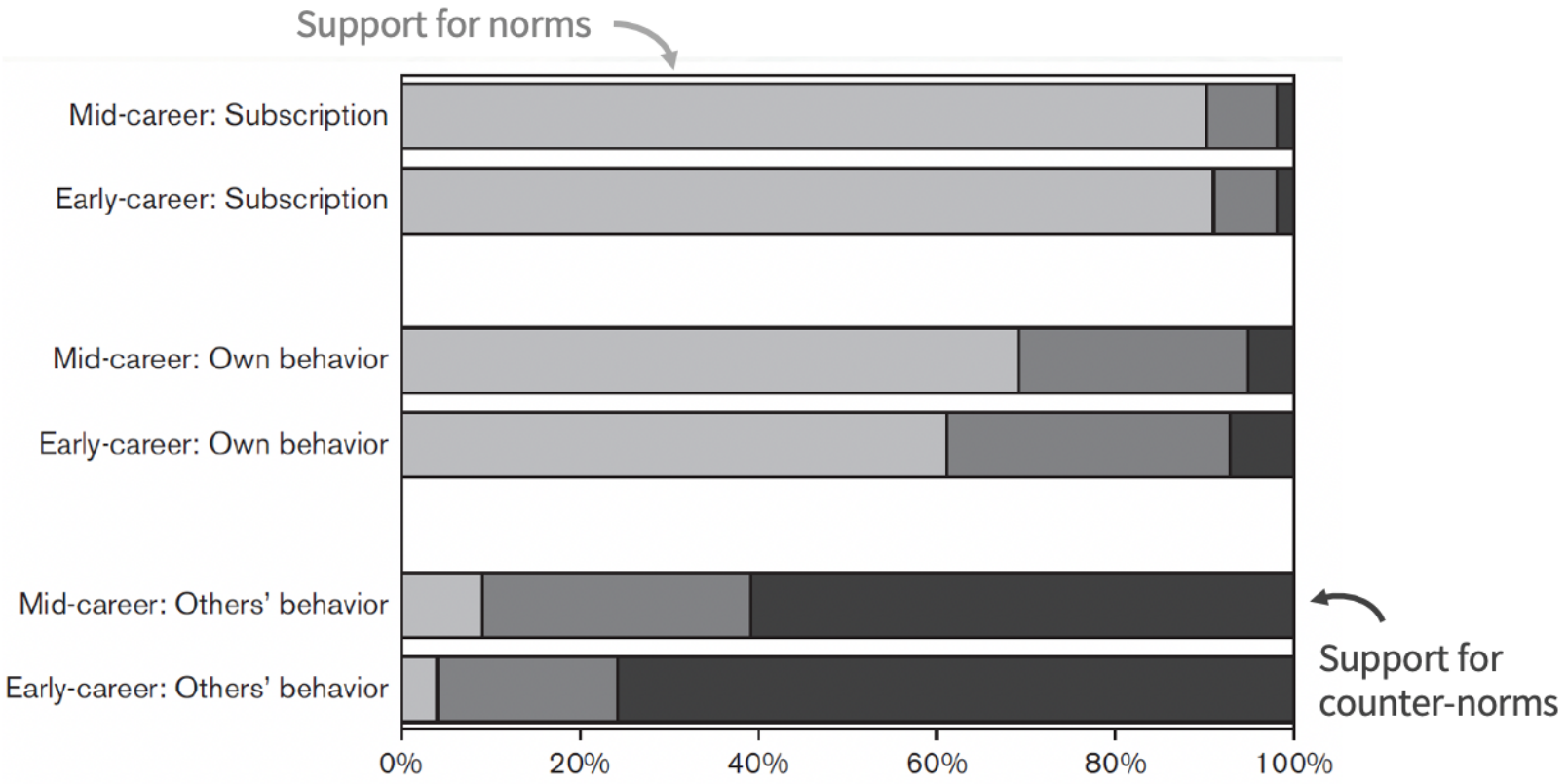
(1) Ask scientists.

Norm support:

"In theory"

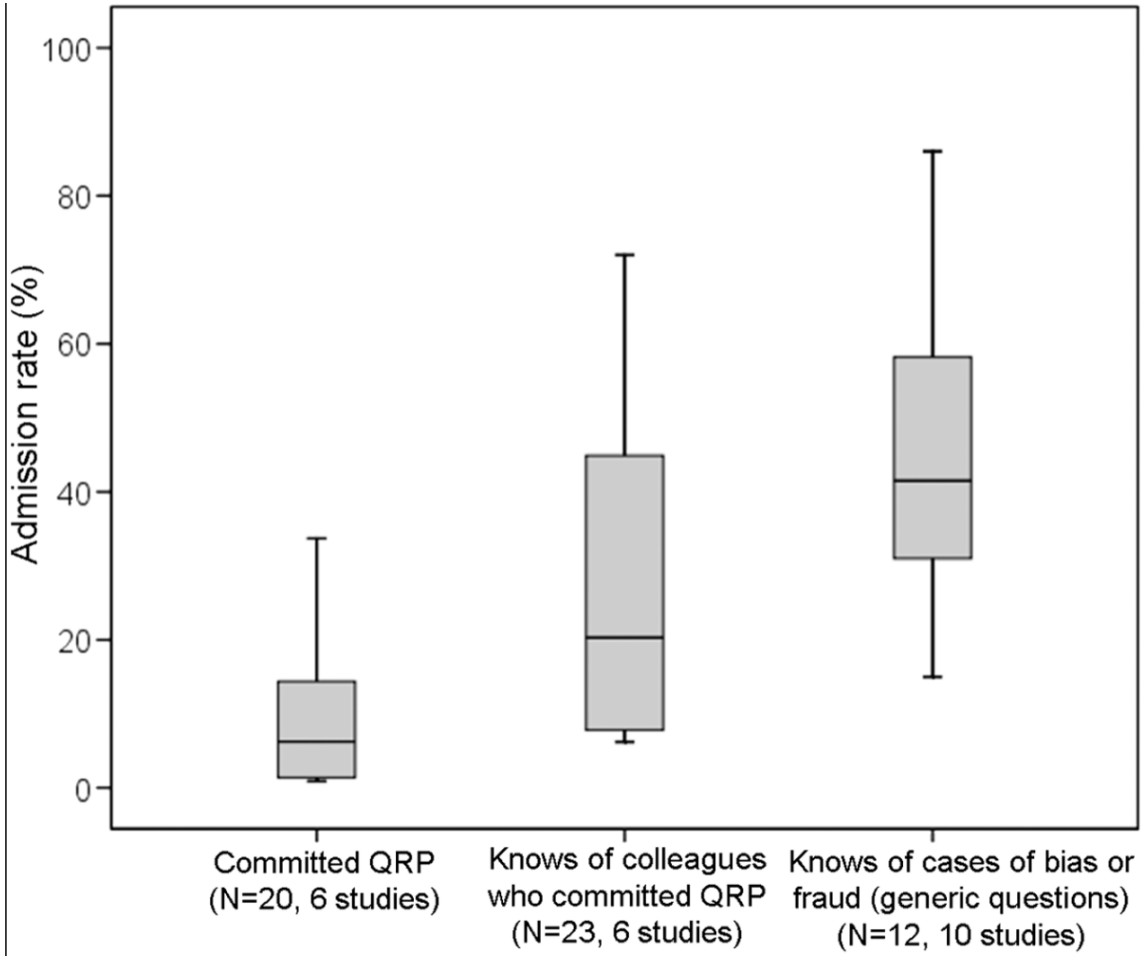
"Me"

"Others"



Christensen et al. (2019) surveyed 3247 US researchers funded by NIH

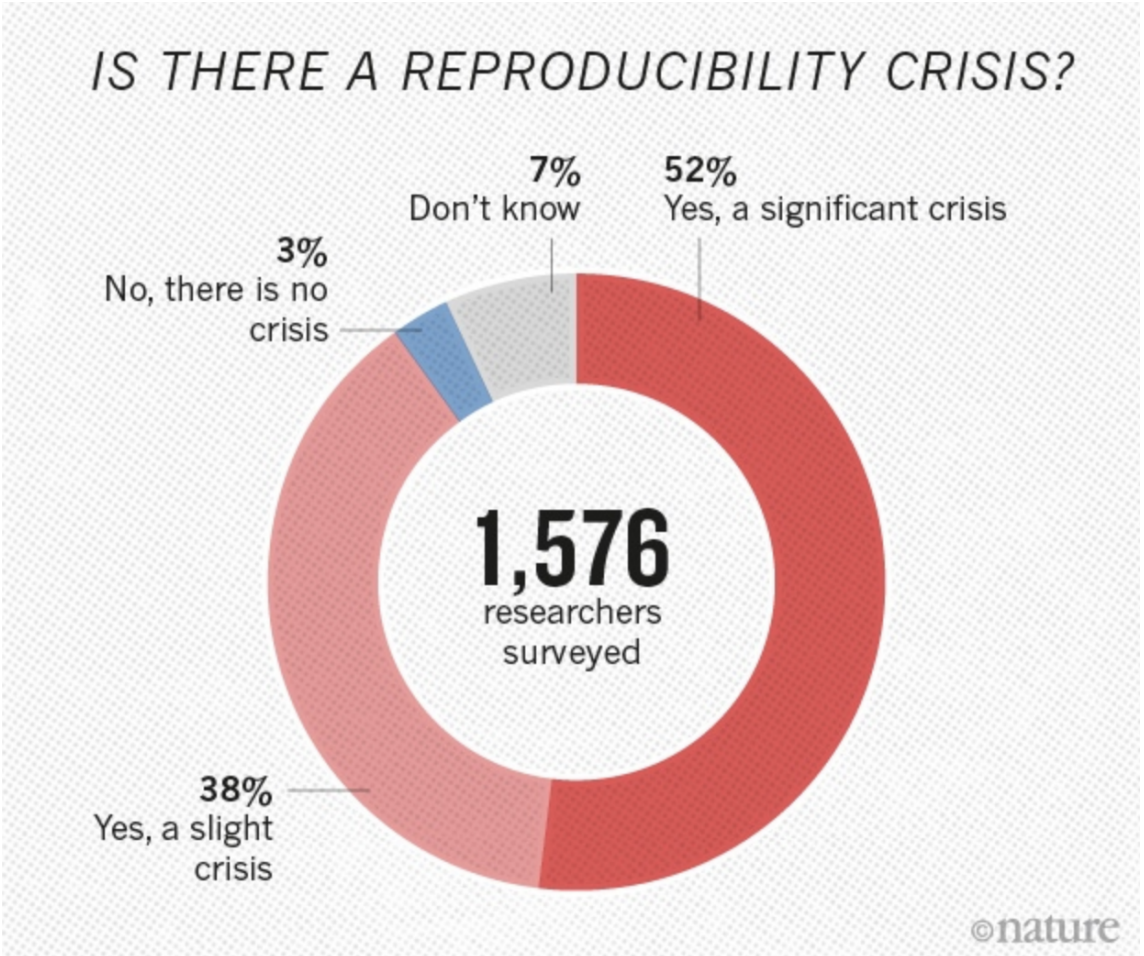
Scientists admit to engaging in questionable research practices.



Fanelli *PLoS ONE* (2011)

Scentists think there is a "reproducibility" crisis

or a "slight" crisis? 🤔



Baker *Nature* (2018)

How do we know that science isn't working?

(2) Look at what we are doing.

Potential sources of **systemic** bias in published research

Deceipt

Fraud, manipulation, fabrication

Significance chasing

P-hacking, publication bias, file drawers, broken peer review

Conflicts of interest

Financial, self-interest, careerism, bad incentives

Integrity Problems

Confirmation bias, dogmatism

> I would like you to really dig into this to find a number of situations or people for which this relationship does hold -- that is where the 1/2 price buffet did result in a difference.

Integrity Problems

Fraud, manipulation, fabrication, etc.

> 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.

Integrity Problems

P-hacking, fishing, data dredging, etc.

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 . . .

Integrity Problems

P-hacking, fishing, data dredging, etc.

> 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 . . .

Integrity Problems

Careerism, attention seeking, least publishable unit, etc.

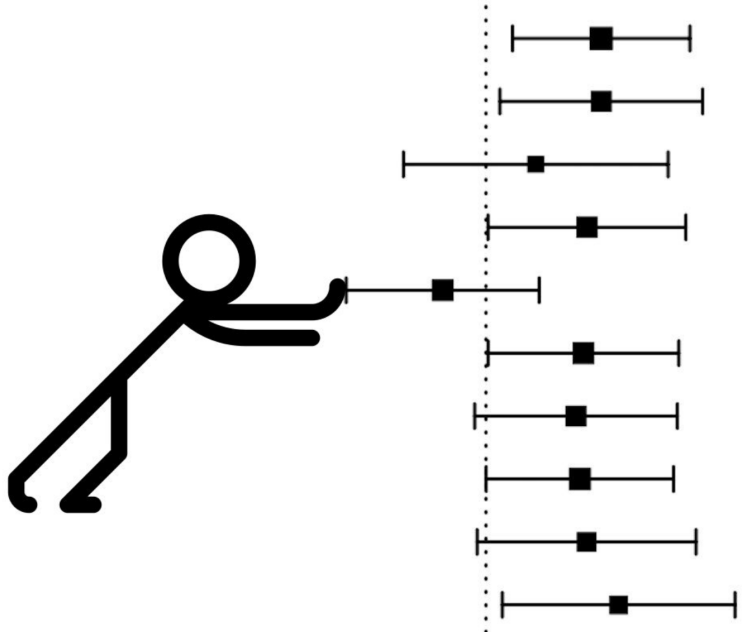
> This is really important to try and find as many things here as possible *before* you come. First, it will make a good impression on people and helps you stand out a bit. Second, it would be the highest likelihood of you getting something publishable out of your visit.

Integrity Problems

Lack of transparency

This entire *research plan* was an email.

A lot of irreproducible or unreliable research stems from Null Hypothesis Significance Testing



Lash (2017), image credit:twitter

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

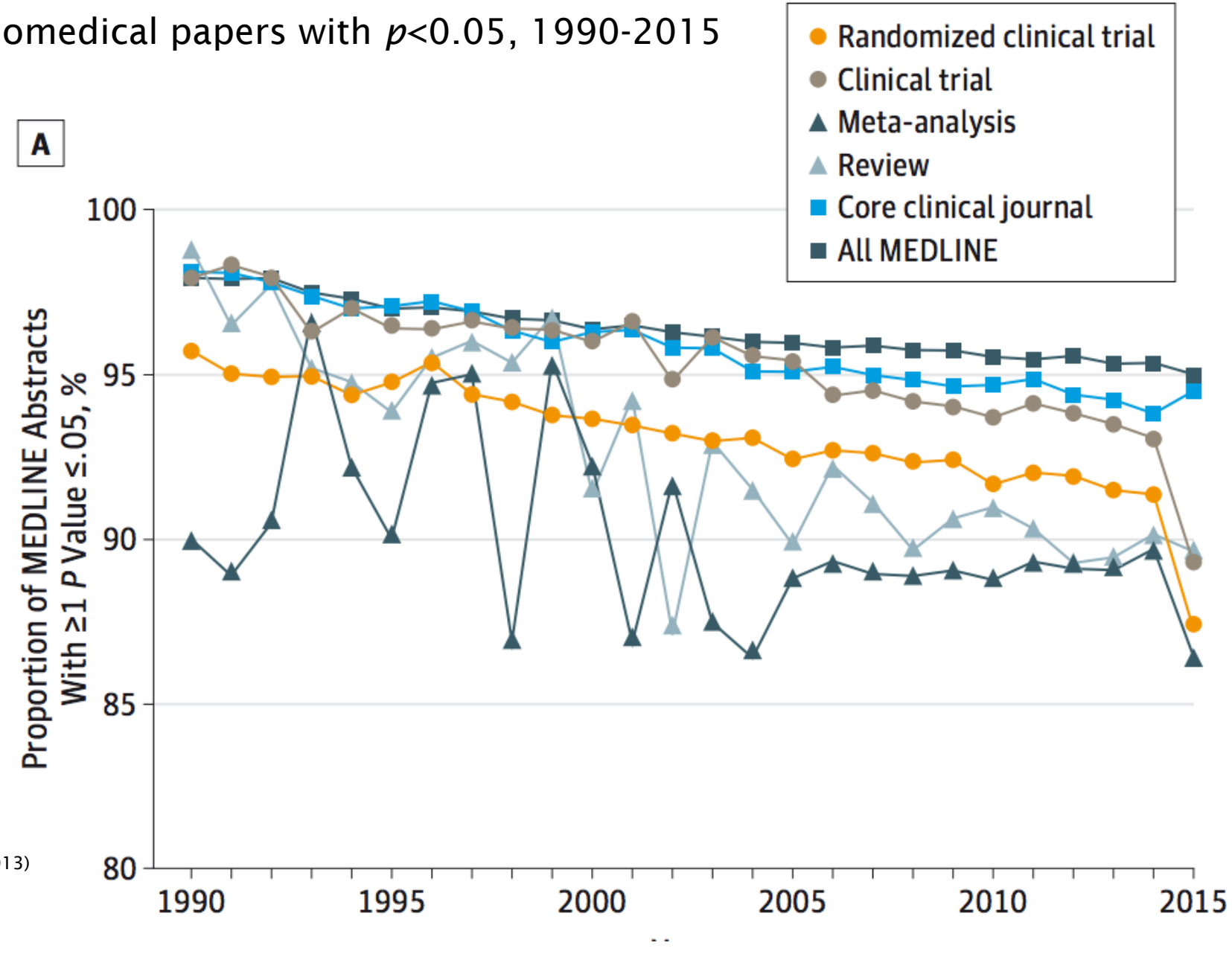
https://www.ahajournals.org/doi/abs/10.1161/jaha.116.004880

How do we know there is p-hacking?

(1) Look at what people are doing.

(2) Everything is significant

90+% of biomedical papers with $p<0.05$, 1990-2015



Chavalarais et al. (2013)

How do we know there is p-hacking?

(1) Look at what people are doing.

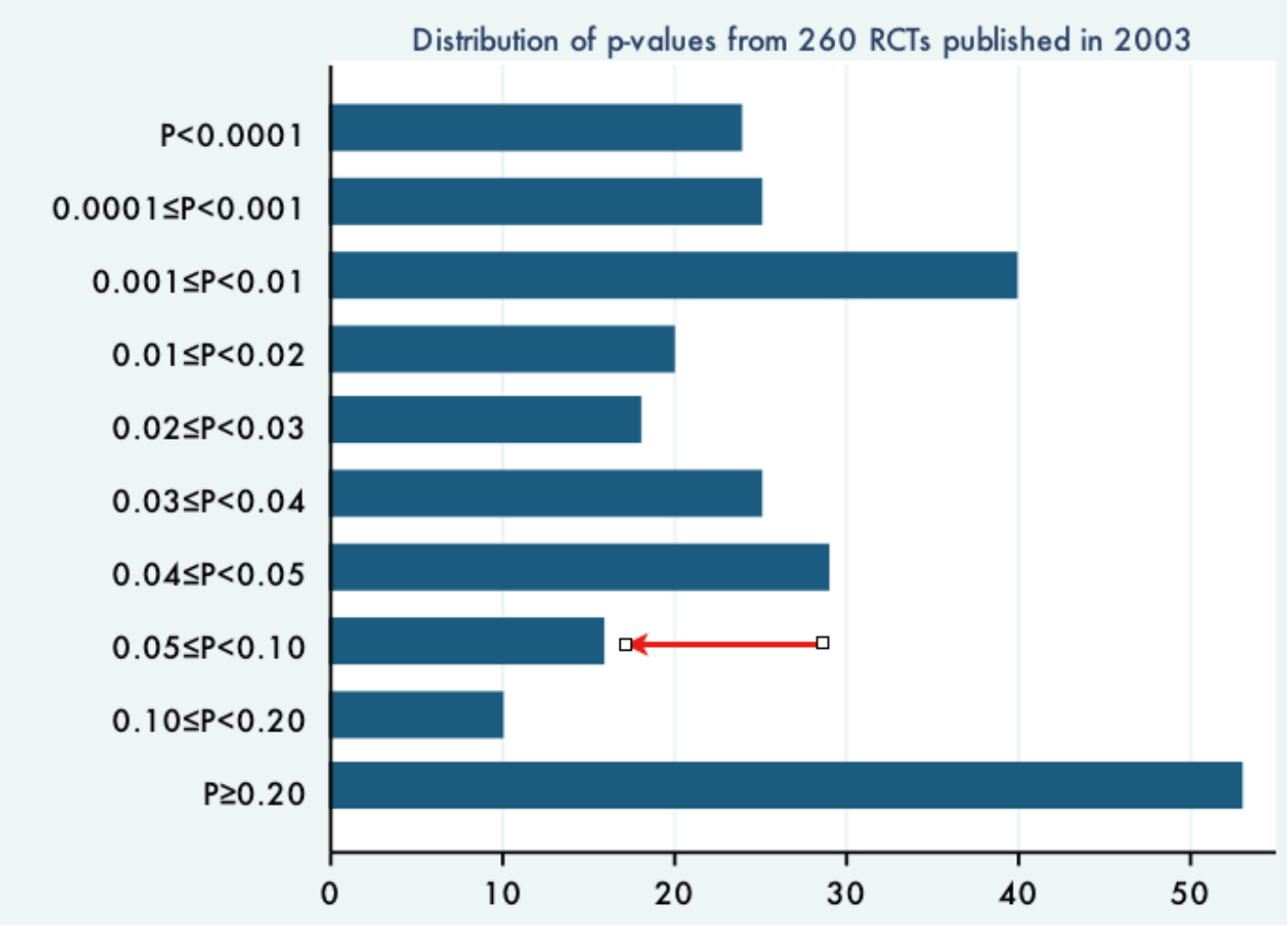
(2) Everything is significant.

(3) Maldistribution of p-values.

Missing p-values just *above* 0.05.

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

P-values from 260 RCTs



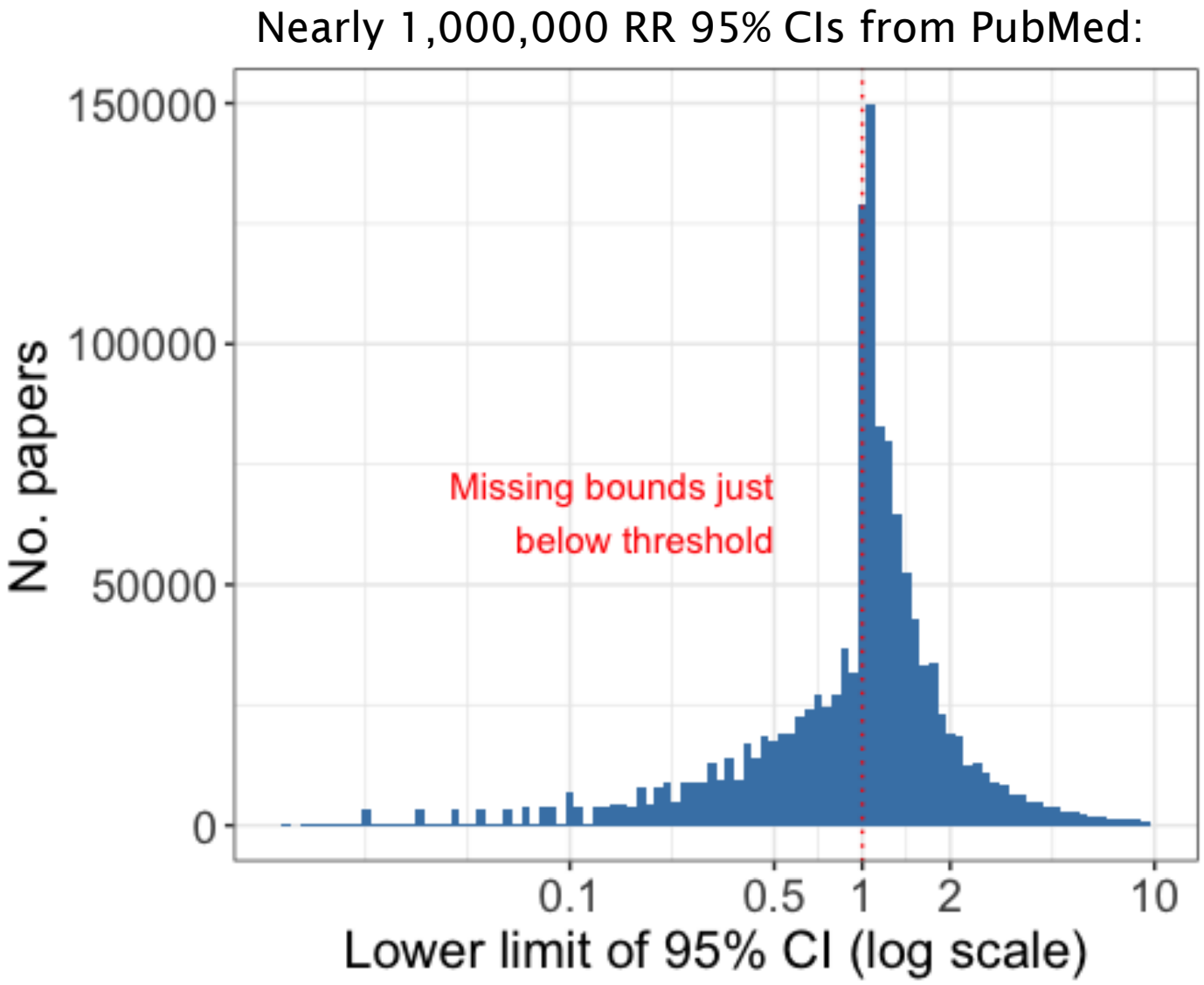
Gotzsche (2006)

Won't 95% confidence intervals help?

No.

Researchers still hack until "significant."

Barnett and Wren (2019)



NHST also leads to missing evidence and publication bias

Missing evidence

Negative studies of antidepressents less likely to be published.

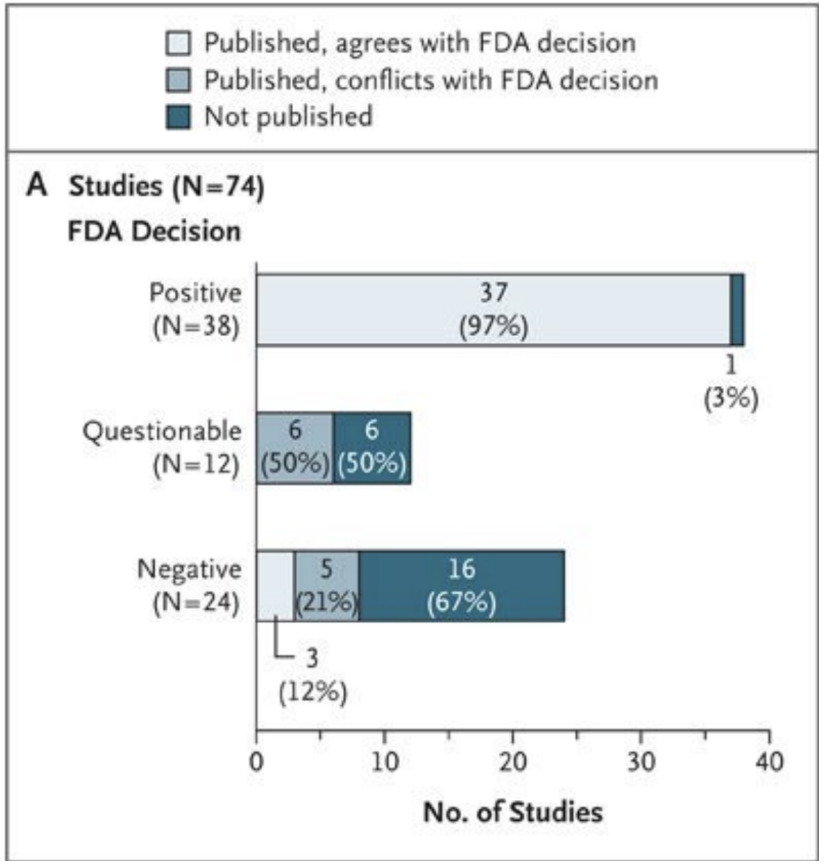
Impacts regulatory decisions.

Caused by researchers, peer review, journal editors, funders.

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. (2008)

Large scale replication efforts find diminished effects.

In Psychology

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.

In Economics

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.

Nosek et al. (2017), Camerer et al. (2016)

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

We have abundant evidence of integrity problems in existing science.

Open science can help

Great. What is it?

Open science aims to provide tools and incentives to make science more transparent, reproducible, and reliable.

How can open science help?

1. Transparent Design

2. Reproducible Workflow

3. Open Dissemination

Design Solutions

Preregistration/pre-analysis plans

Reporting guidelines

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

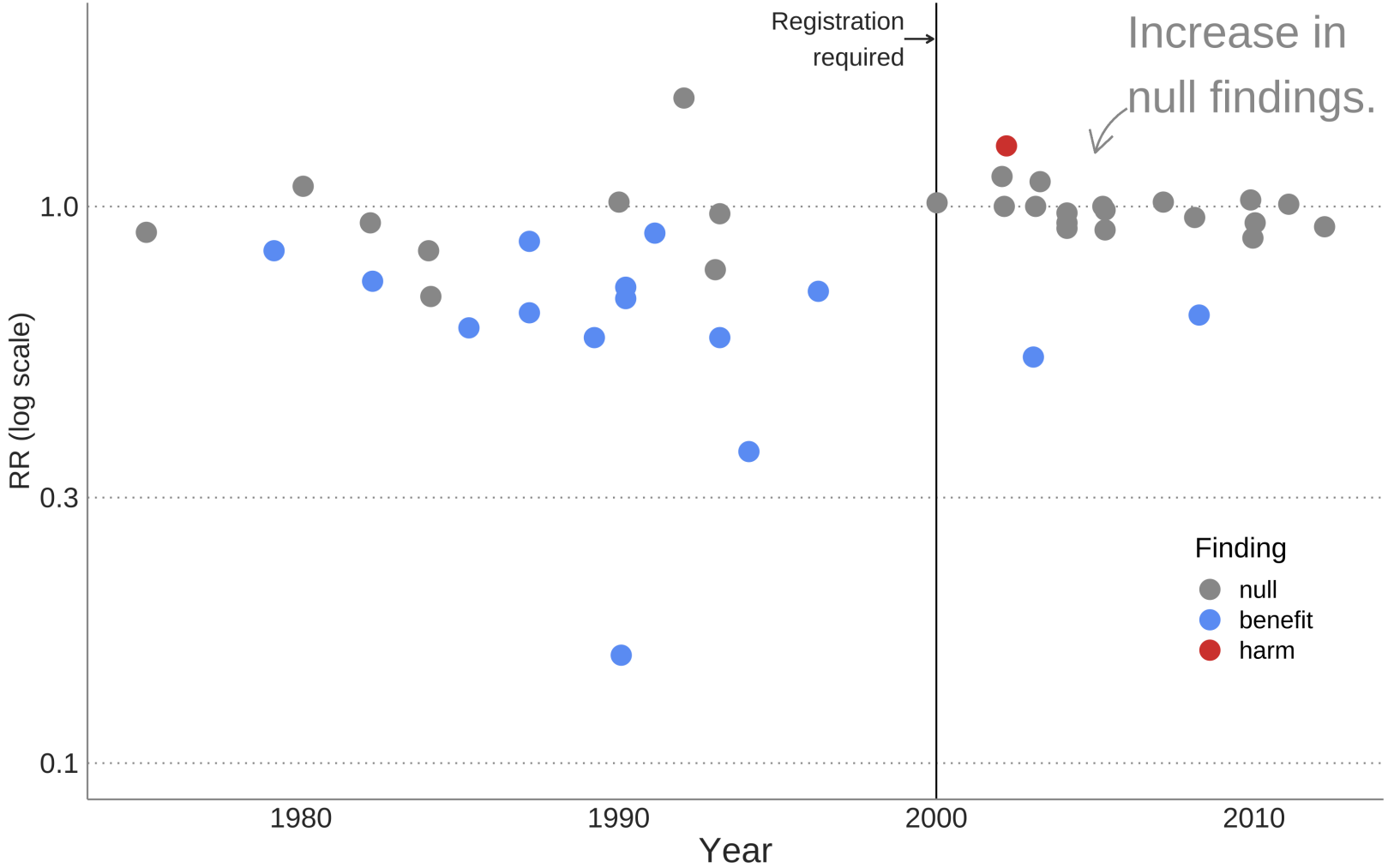
Robust to fieldwork, data snooping.

What is 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.
- Not a straightjacket. Transparent deviations acceptable.

Registration is useful

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



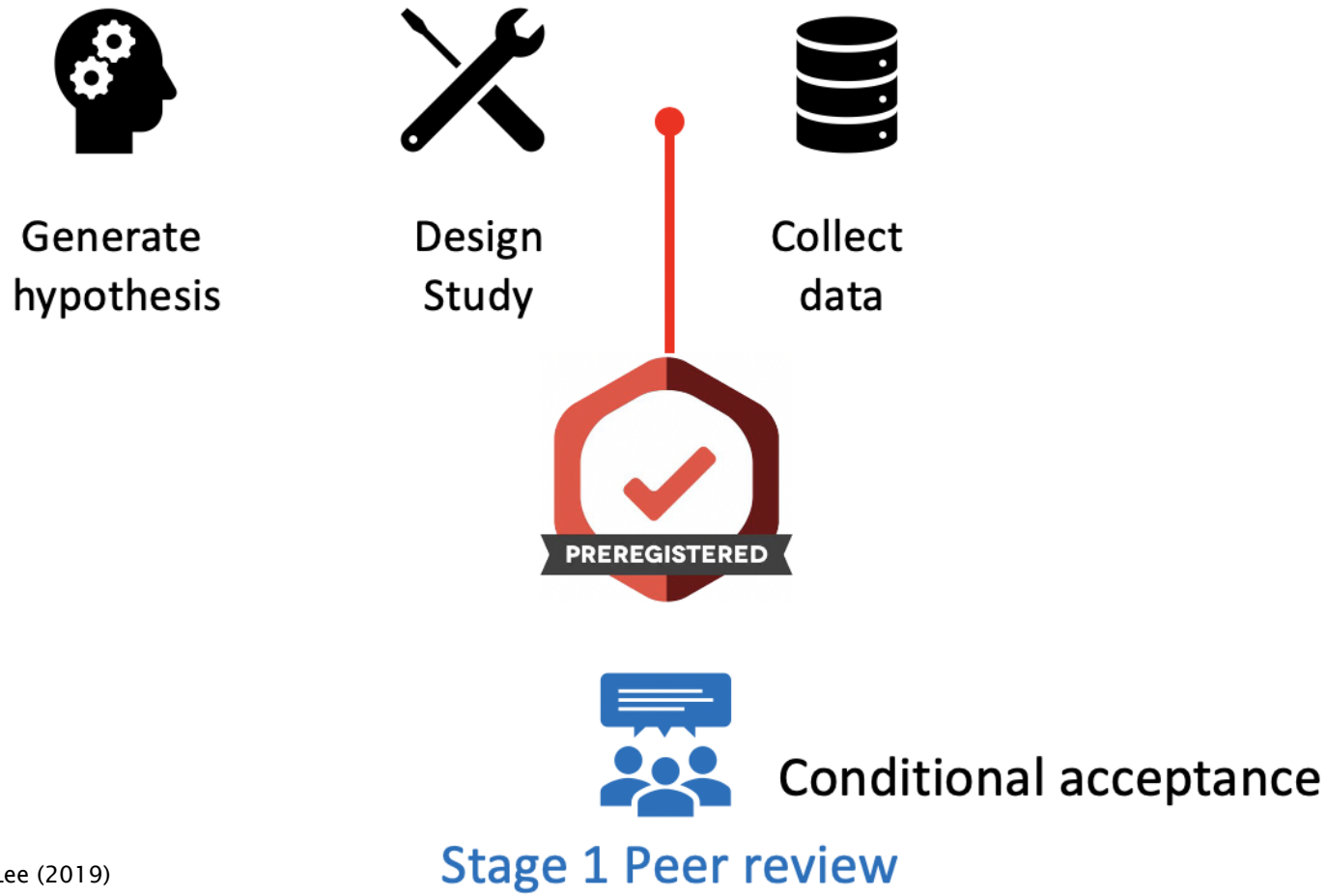
redrawn from Kaplan and Irwin (2015)

What if my results are null?

You showed us that they won't get published!

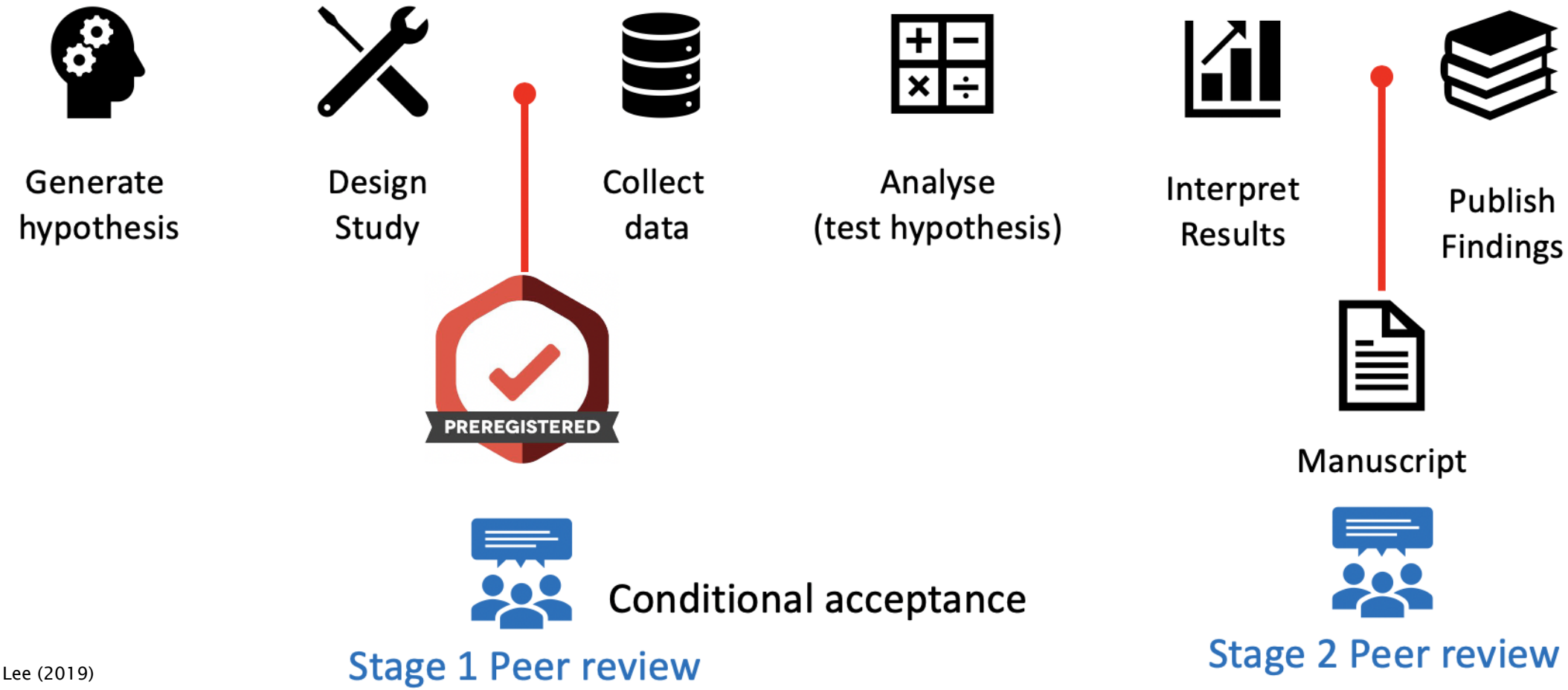
I have to make rent, you know.

Emphasis on design: Registered Reports



Lee (2019)

Emphasis on design: Registered Reports

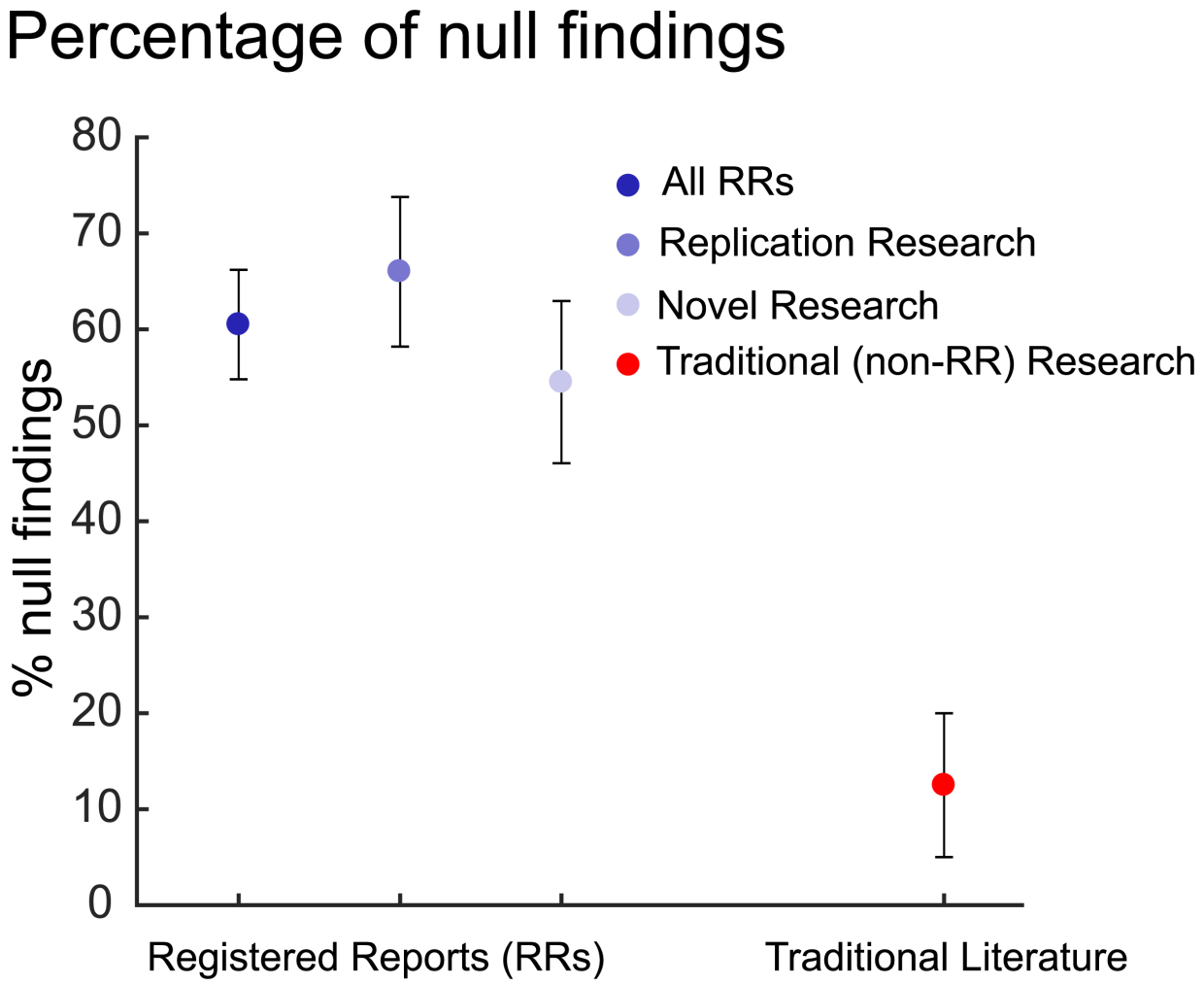


Lee (2019)

RRs in Psychology

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

Big difference from non-registered studies.

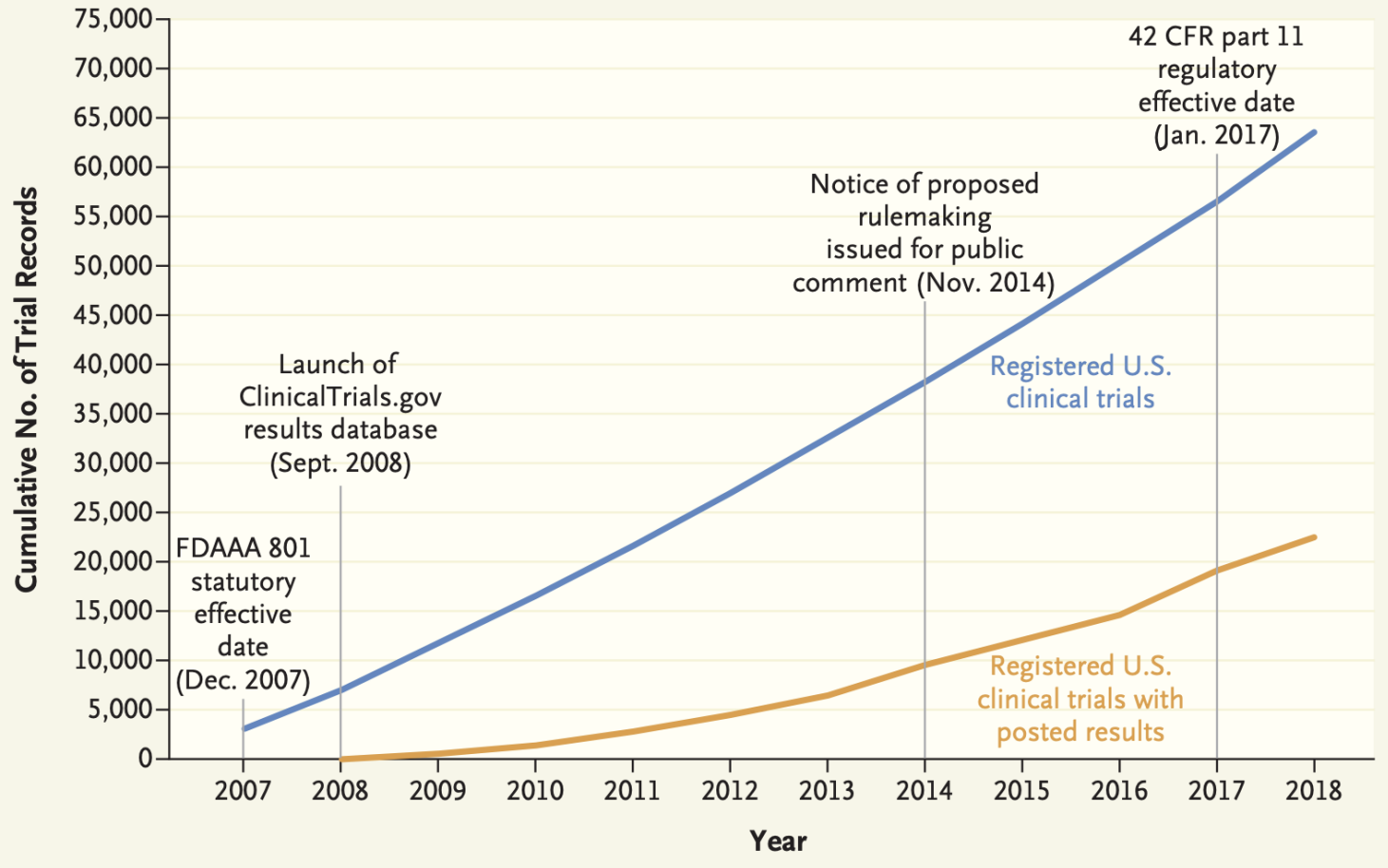


Allen & Mehler, *PLoS Biology* (2019)

Registration is useful

but not sufficient

A majority of registerd 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)

Analytic Solutions

Workflow Management

Documentation

Literate Programming

Version Control

Dynamic Documents

Tools have consequences

SEPT2 gene

👇

👇

2-Sep

Boddy (2016), Ziemann (2016)

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl ver-

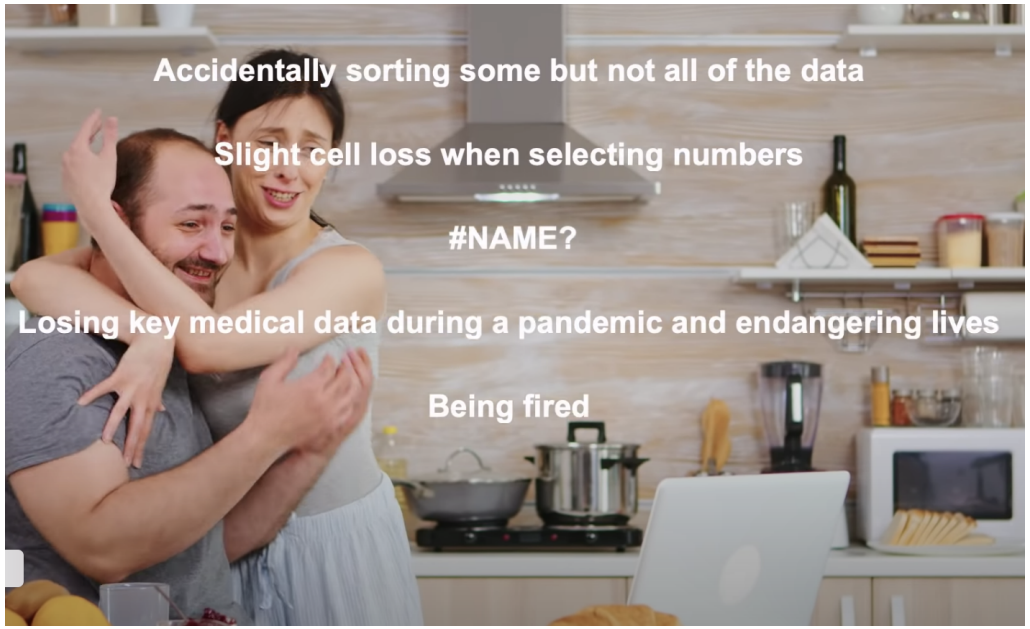
More recently...

Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

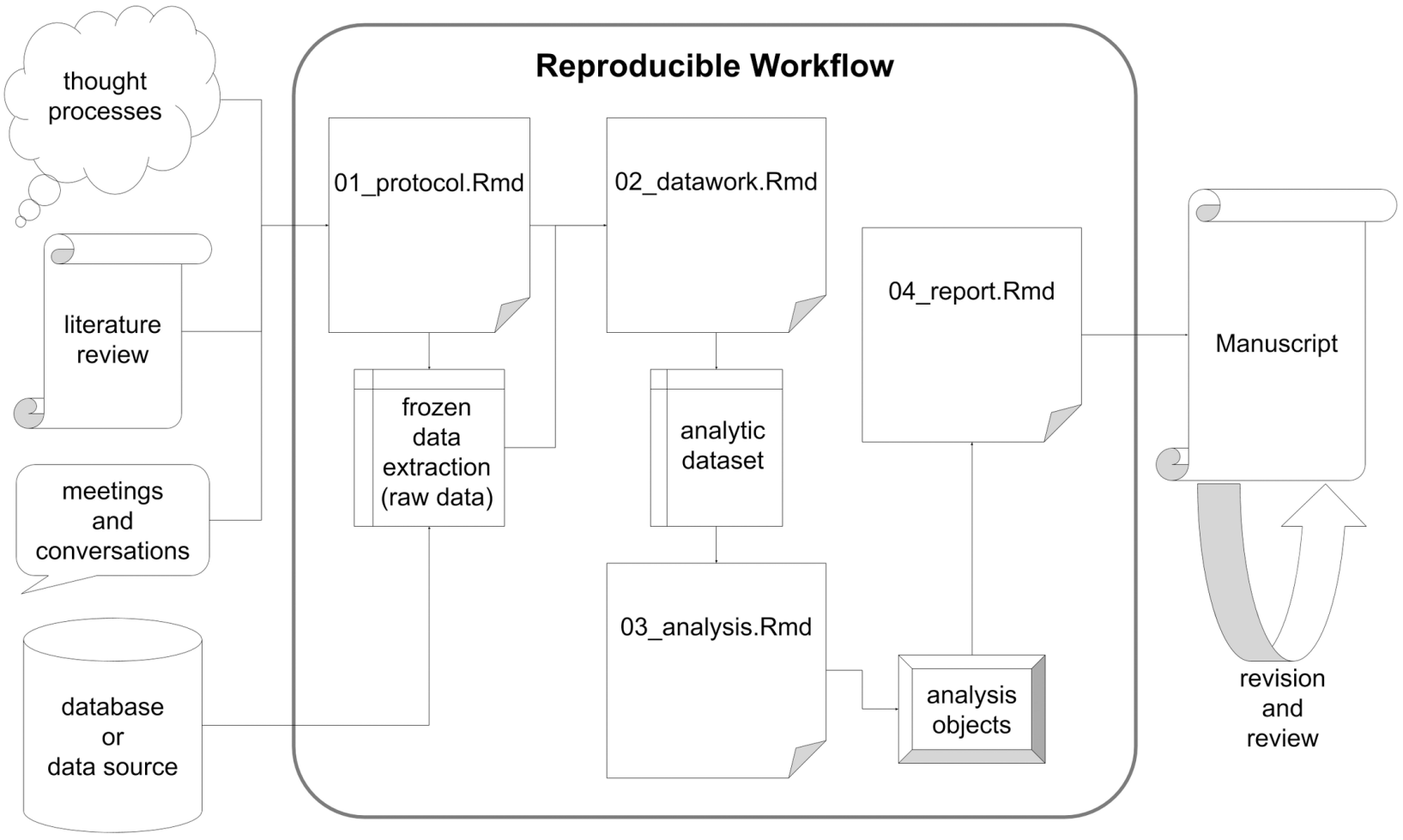
Sources: The Guardian (2020-10-06), YouTube

Are Spreadsheets® right for you? Side effects may include:



Modern workflow

- Code and data separated.
- Raw data never altered.
- Figures and tables created by scripts.
- No copy/paste
- Results + text in a single *dynamic* document.



Krieger et al. PLoS ONE (2019)

Dissemination Solutions

Replication Files

Sharing

Replication files provide the 'recipe' for reproducing your results.

Facilitate reproducibility

Anyone can reproduce your tables and figures.

Detects errors

Coding is hard. We all make mistakes.

Extends work

Probes reliability of findings, answers new questions.

The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather

By Olivier Deschênes and Michael Greenstone*

The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather (pp. 354-385)

Olivier Deschênes and Michael Greenstone

Abstract/Tools | Full-text Article | Download Data Set | Link to Appendix

percent. This estimate is robust to numerous specification checks and relatively precise, so large negative or positive effects are unlikely. We also find the hedonic approach—which is the standard in the previous literature—to be unreliable because it produces estimates that are extremely sensitive to seemingly minor choices about control variables, sample, and weighting. (*JEL* L25, Q12, Q51, Q54)

The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment†

By Anthony C. Fisher, W. Michael Hanemann, Michael J. Roberts, and Wolfram Schlenker*

Fisher et al. found:

1. data and coding errors in DG's weather data, agricultural data, and the construction of climate-change scenarios;
2. the particular climate change scenario which is used for impact predictions; and
3. standard errors that are biased due to spatial correlation.

"Correcting DG's data and coding errors makes predictions for climate-change impacts unambiguously negative in all but one specification."

Am Econ Rev (2012)

The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Reply

By Olivier Deschênes and Michael Greenstone*

Fisher et al. (2012) (hereafter, FHRS) have uncovered coding and data errors in our paper, Deschênes and Greenstone (2007) (hereafter, DG). We acknowledge and are embarrassed by these mistakes. We are grateful to FHRS for uncovering them. We hope that this Reply will also contribute to advancing the literature on the vital question of the impact of climate change on the US agricultural sector.

Am Econ Rev (2012)

Why share?

Credibility

Others can reproduce or interrogate your findings.

Social Good

Resource for other questions and new ideas.

Changing norms

Professional norms are insufficient to change behavior.

Rationale for sharing data and code

Online repositories last longer, are indexed.

Concerns:

- Can usually be embargoed, or provide only what is necessary for replication (e.g., unused survey Qs).
- Biggest risk isn't having your data/ideas stolen, it's having your research ignored! (King 1995)
- *More* difficult if research products are proprietary.

Many resources to help

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 80–88
https://doi.org/10.1080/00031305.2017.1375986

Packaging Data Analytical Work Reproducibly Using R (and Friends)

Ben Marwick[a], Carl Boettiger[b], and Lincoln Mullen[c]

[a]University of Washington, Seattle, WA; [b]University of Wollongong, Wollongong, New South Wales; [c]University of California, Berkeley, CA; [d]George Mason University, Fairfax, VA

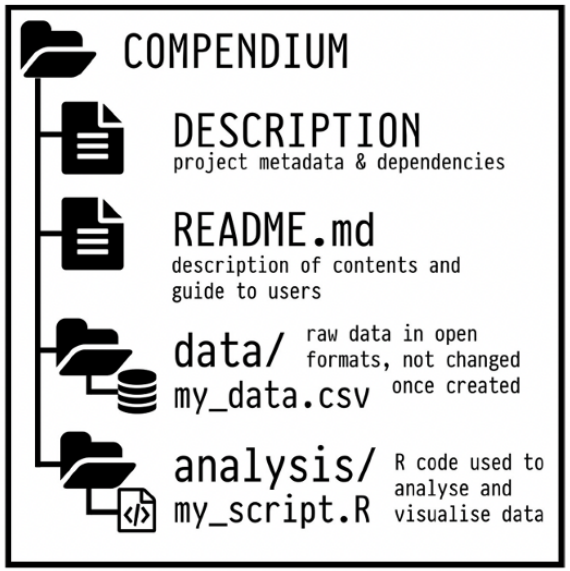
ABSTRACT
Computers are a central tool in the research process, enabling complex and large-scale data analysis. As computer-based research has increased in complexity, so have the challenges of ensuring that this research is reproducible. To address this challenge, we review the concept of the research compendium as a solution for providing a standard and easily recognizable way for organizing the digital materials of a research project to enable other researchers to inspect, reproduce, and extend the research. We investigate how the structure and tooling of software packages of the R programming language are being used to produce research compendia in a variety of disciplines. We also describe how software engineering tools and services are being used by researchers to streamline working with research compendia. Using real-world examples, we show how researchers can improve the reproducibility of their work using research compendia based on R packages and related tools.

ARTICLE HISTORY
Received May 2017
Revised August 2017

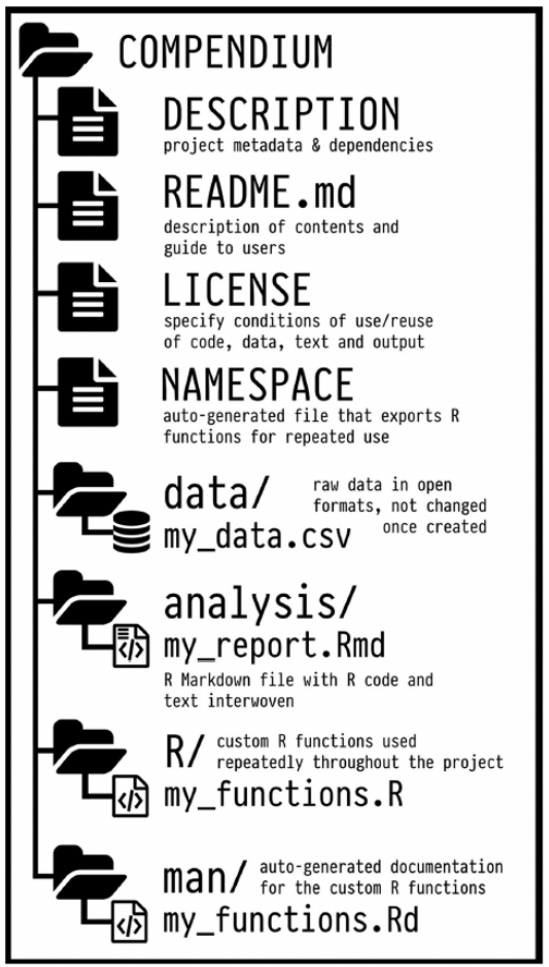
KEYWORDS
Computational science; Data science; Open source software; Reproducible research

Can be done for any size project 👇

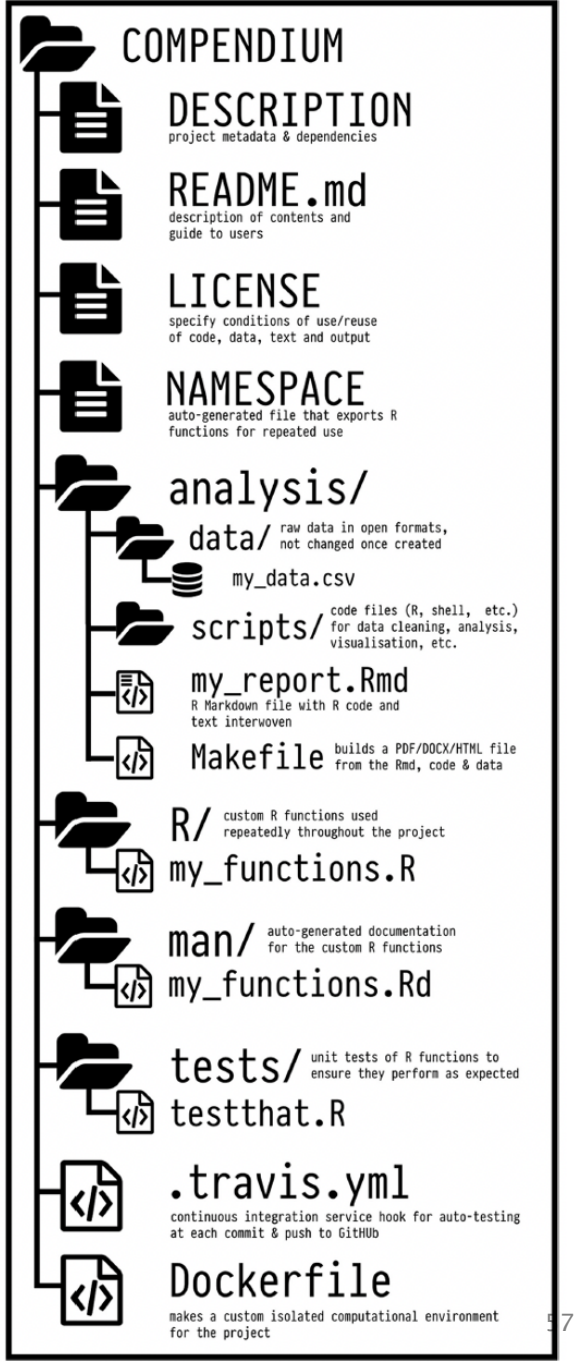
Small



Medium

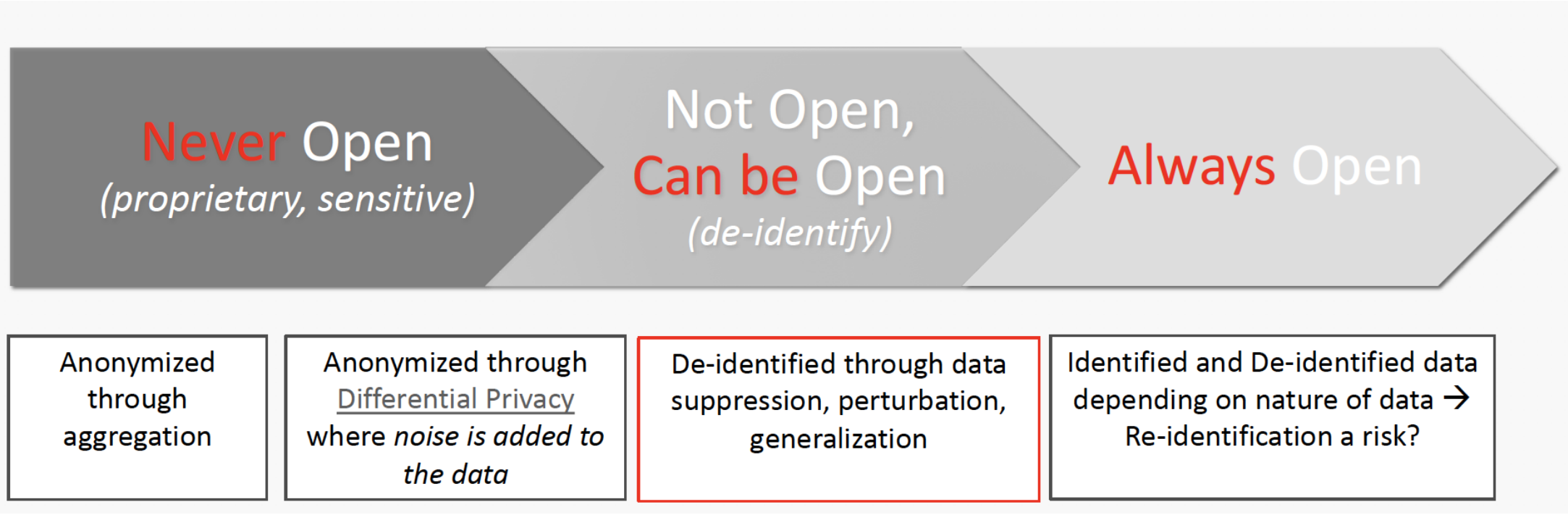


Large



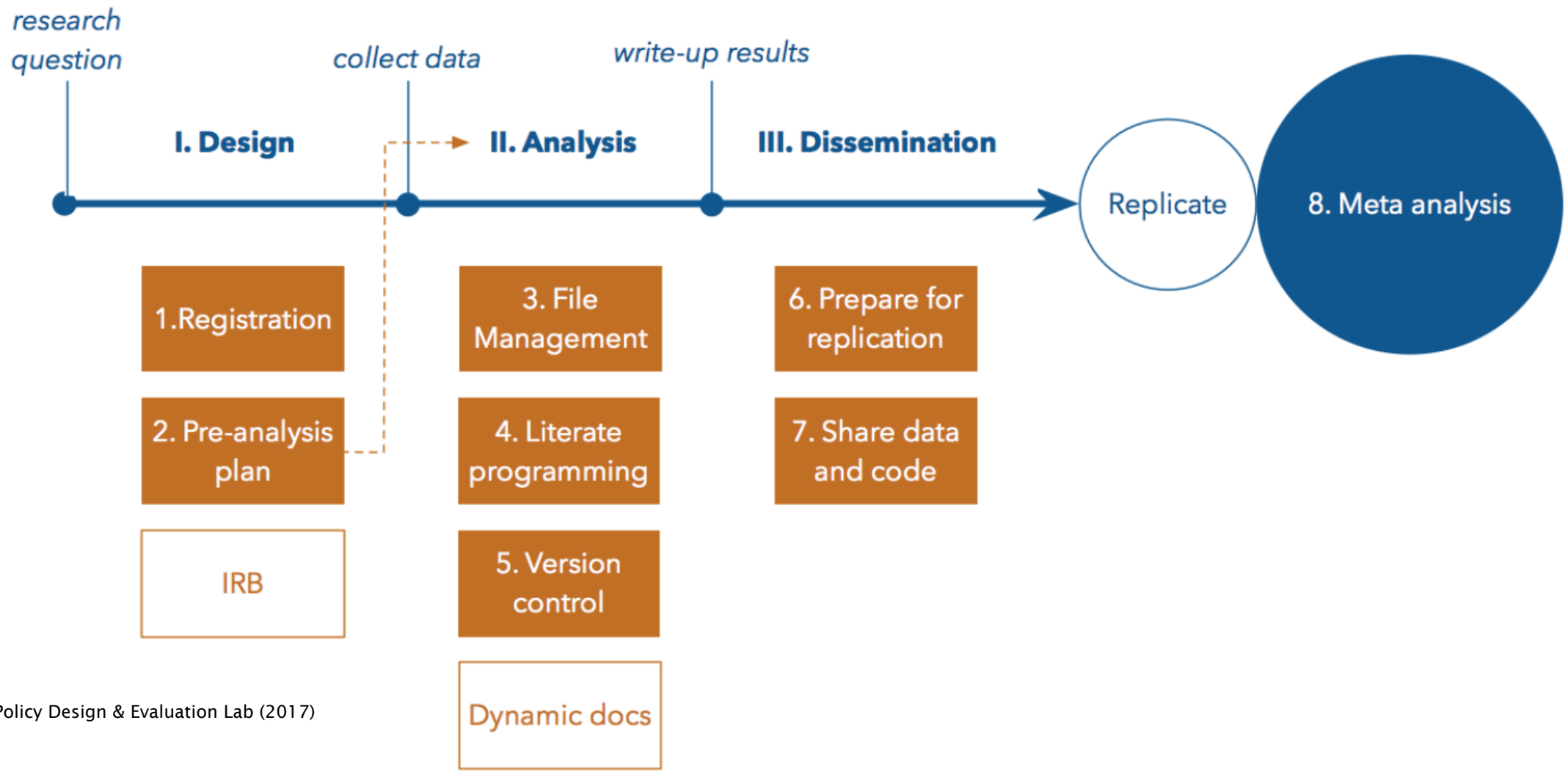
Not everything can (or should) be shared

Spectrum for sharing sensitive material:



Source: Jennifer Sturdy (https://osf.io/5yq4u/)

A reproducible path forward: Reminaging the research lifecycle?



Policy Design & Evaluation Lab (2017)

Challenges

Open science adds work

Competes with incentives, career, and time demands

Open materials are open to all

Biases still present (see concerns about EPA rulemaking changes)

Open science $\neq$ true science

Only a minimum standard for transparency.

Preregistered, planned studies with completely transparent methods and open protocols, data, and code may be bad

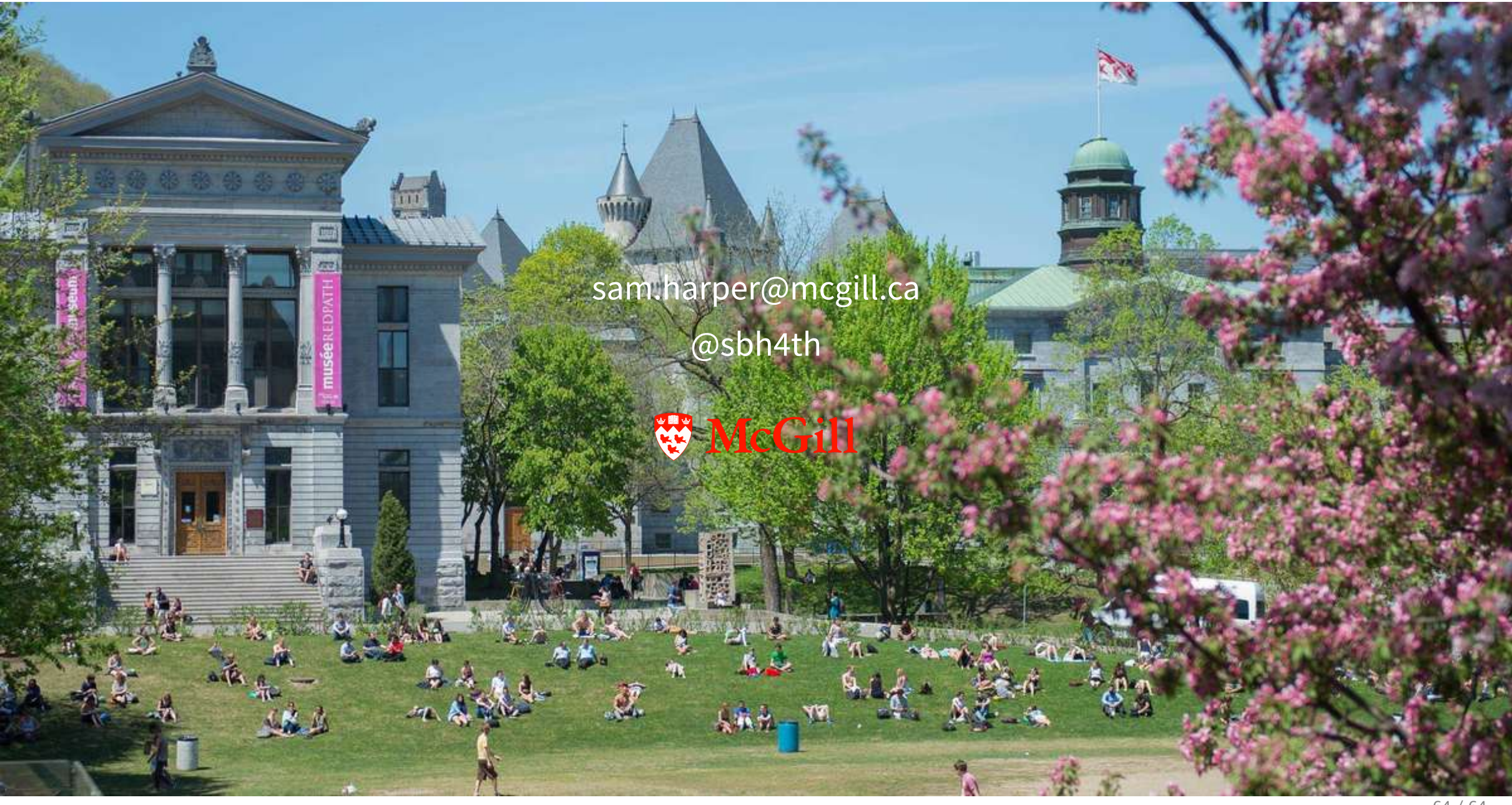
Open Science Does Not Mean True Science

Summary points: Why Embrace Open Science?

1. To contribute to improving scientific integrity.

2. To move professional norms in a positive direction.

3. To improve the quality of your own reserach.


musée REDPATH
sam.harper@mcgill.ca
@sbh4th
McGill