Reproducible Research: Why and How

SER Pre-Conference Workshop

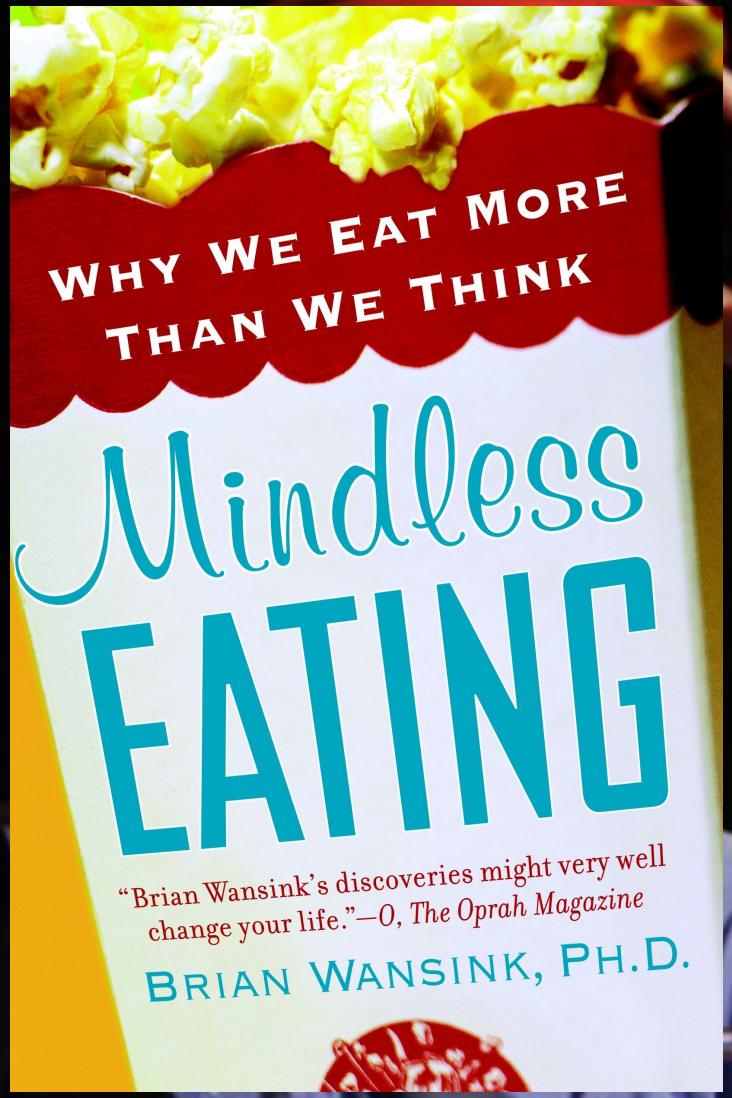
Sam Harper

2020-10-30

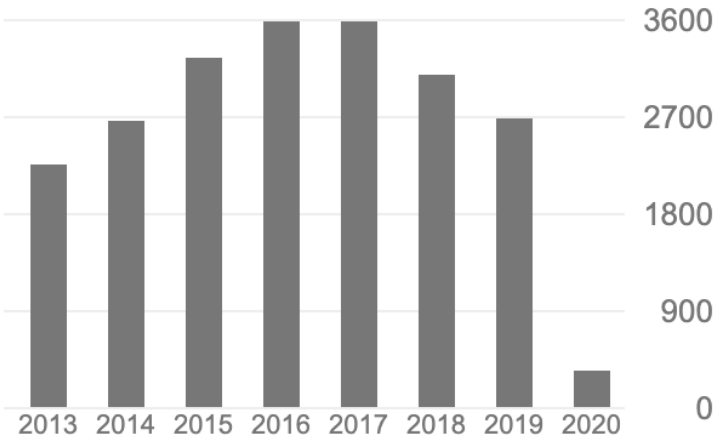
I am a social epidemiologist at McGill University.

I work mainly on evaluating programs and policies on social inequalities in health.

I have nothing to disclose, other than a strong commitment to open science



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.

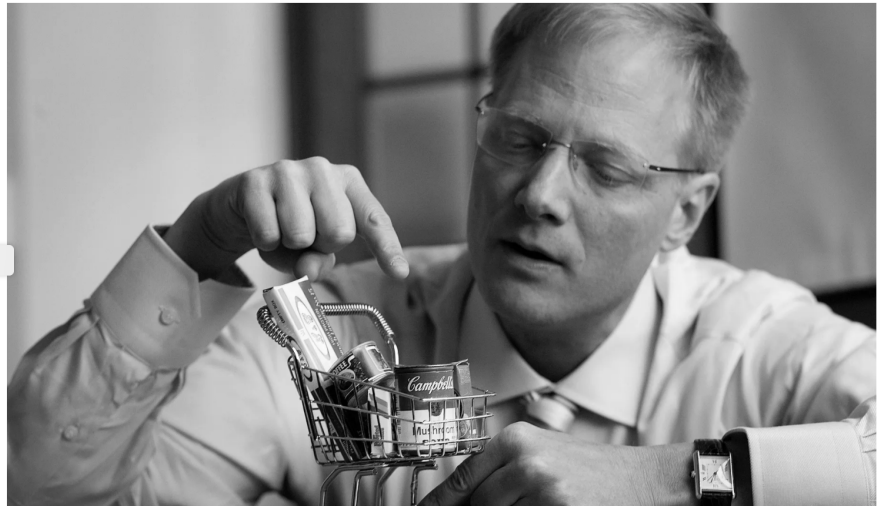
Archived post here.

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.

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-

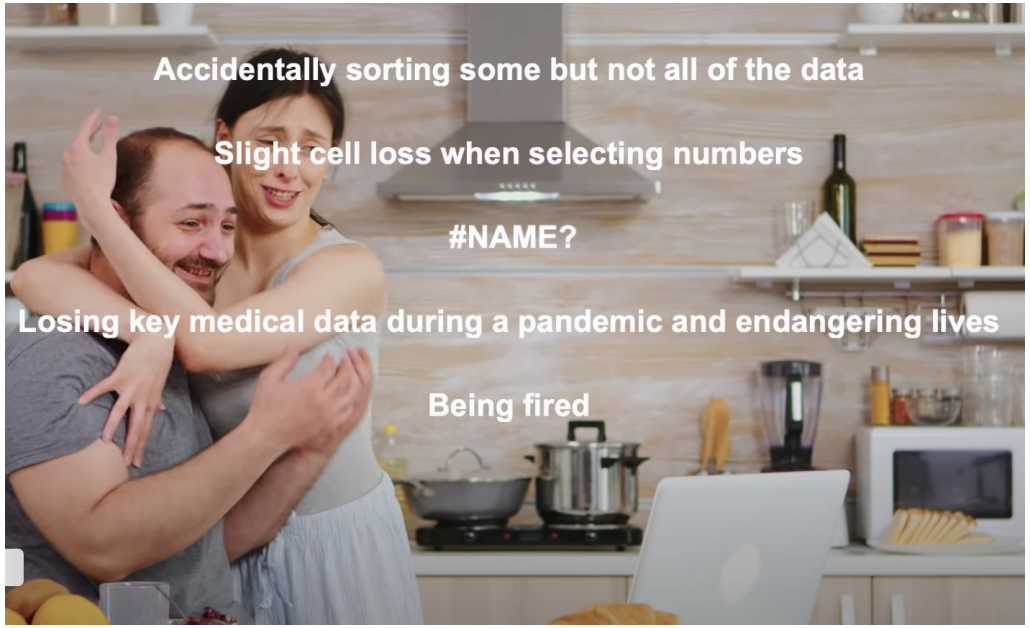
More recently...

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:



The integrity of science is compromised by non-reproducible research.

There are tools to help you.

Setting expectations

Today is not about:

- Mastering software
- Learning to code
- Mastering version control
- Mastering statistical analysis

Today is about:

- *Why* to do reproducible research.
- Understanding concepts of *how* to do it.
- Getting familiar with tools to help.
- Learning where to find out more.

What is the plan?

Plan for today

1. Scientific Integrity Problems (1220h-1250h) ☕
2. Design Solutions (1300h-1330h)
3. Analytic Solutions (1330h-1350h, ☕ 1400h-1450h ☕)
4. Dissemination Solutions (1500h-1530)
5. Reproducible Example (1530h-1600h)

Code of Conduct

Do

- Be respectful.
- Ask questions in the chat.
- Use the 'raise your hand' feature to ask a question or make a comment.
- Interrupt me if I didn't notice your chat or 'hand'.
- Feel free to turn your camera on (if you are comfortable).

Don't

- Worry about taking notes (but feel free to do so). You will have access to all of the material for the workshop when we are finished.
- Be disrespectful or rude.

Let's go!

Reproducible Research: Why and How

Part 1: Integrity Problems

Sam Harper

SER Pre-Conference Workshop
2020-10-30

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

Mertonian Norms in Science

Core Values of Scientific Research

1. Universalism
2. Communality
3. Disinterestedness
4. Organized Skepticism

A NOTE ON SCIENCE AND DEMOCRACY

by ROBERT K. MERTON

SCIENCE, as any other activity involving social collaboration, is subject to shifting fortunes. Difficult as the very notion may appear to those reared in a culture which grants science a prominent if not a commanding place in the scheme of things, it is evident that science is not immune from attack, restraint and repression. Writing a scant thirty-five years ago, Veblen could observe that the faith of western culture in science was unbounded, unquestioned, unrivalled. The revolt from science which then appeared so improbable as to concern only the timid academician who would ponder all contingencies, however remote, has now been forced upon the attention of scientist and layman alike. Local contagions of anti-intellectualism threaten to become epidemic.

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

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, hypotheses, theories, and innovations, even those that challenge or contradict 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, resisting alternatives.

Potential sources of "bias" in published research

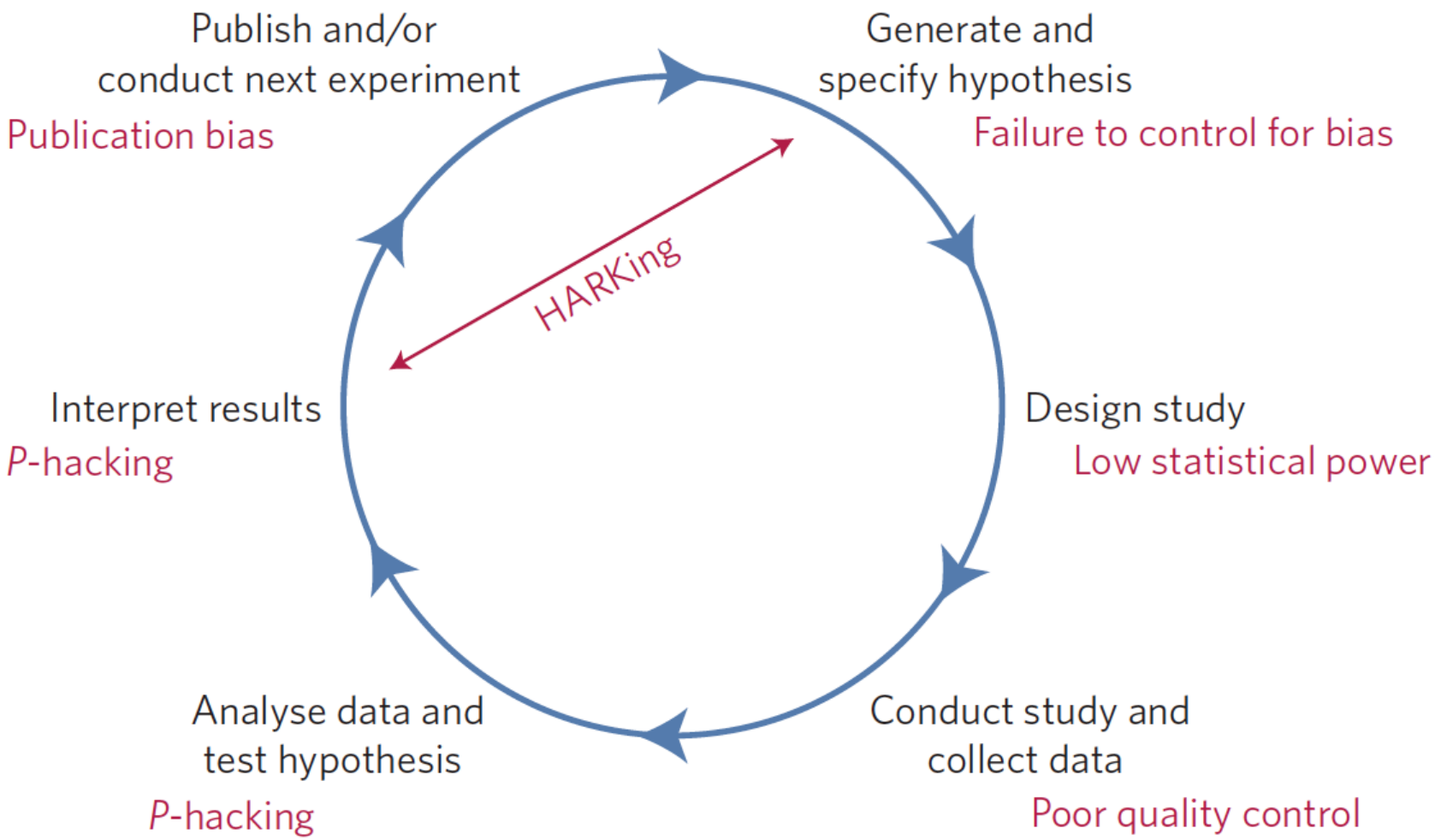
Usual explanations

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.

Affects the entire research lifecycle



Munafo et al. (2017)

How do we know that science isn't working?

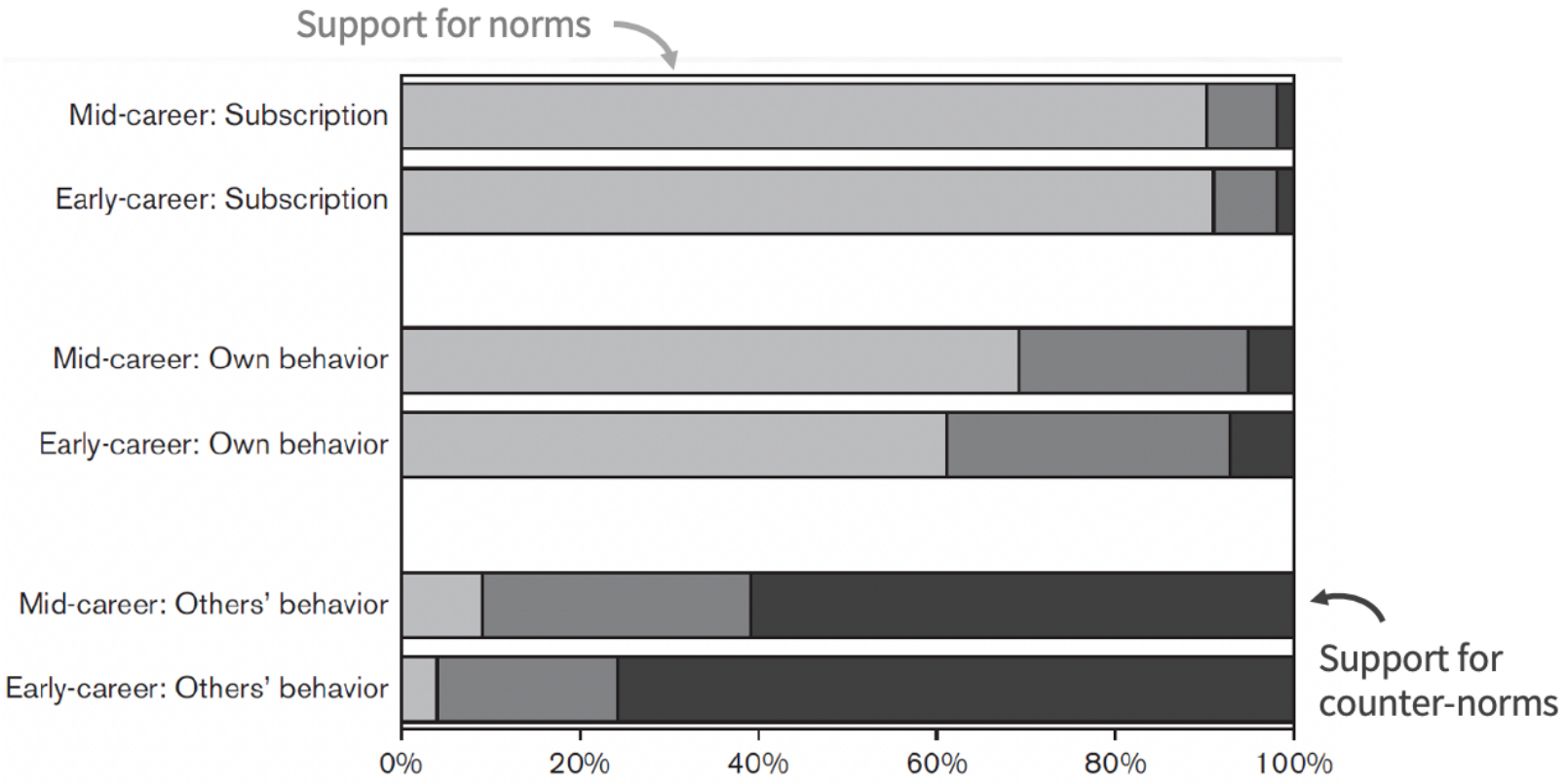
Ask scientists.

Norm support:

"In theory"

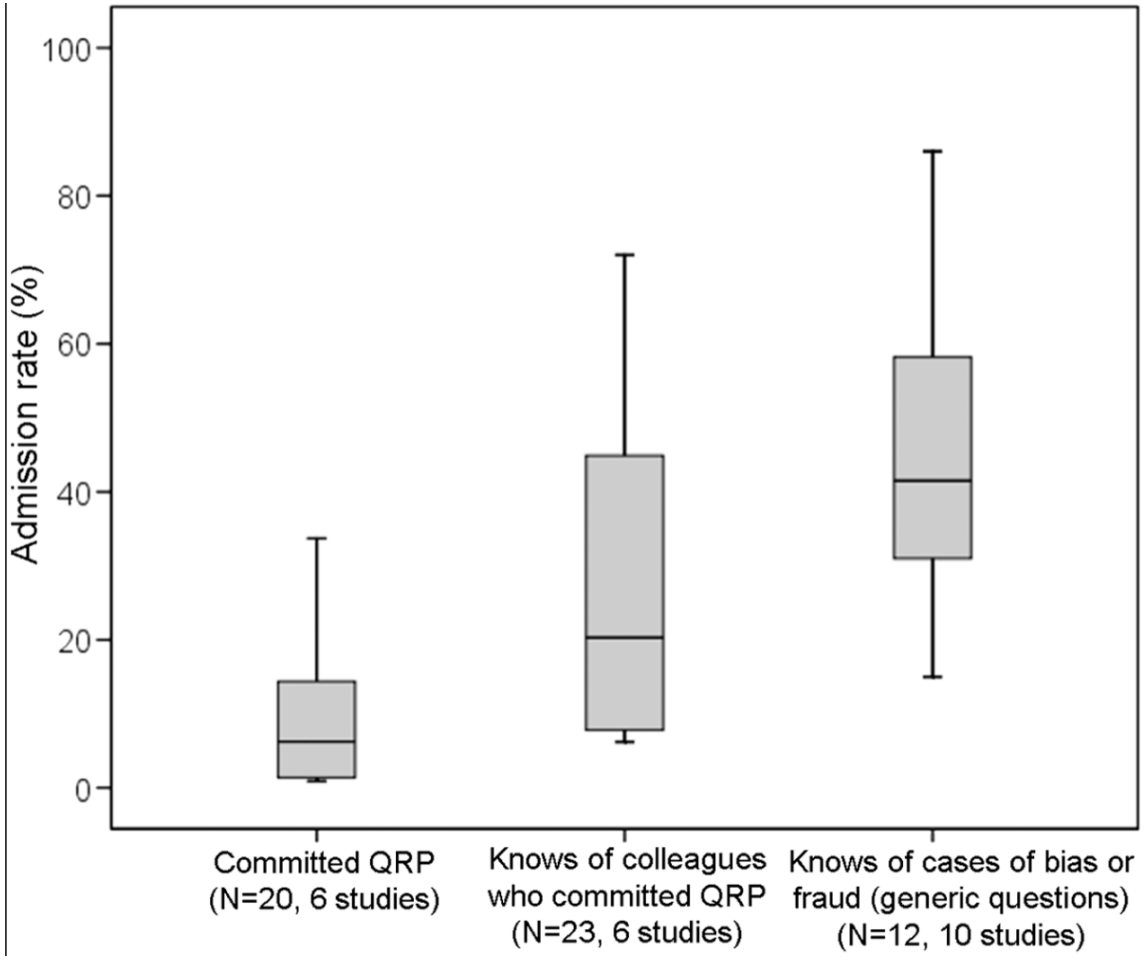
"Me"

"Others"



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

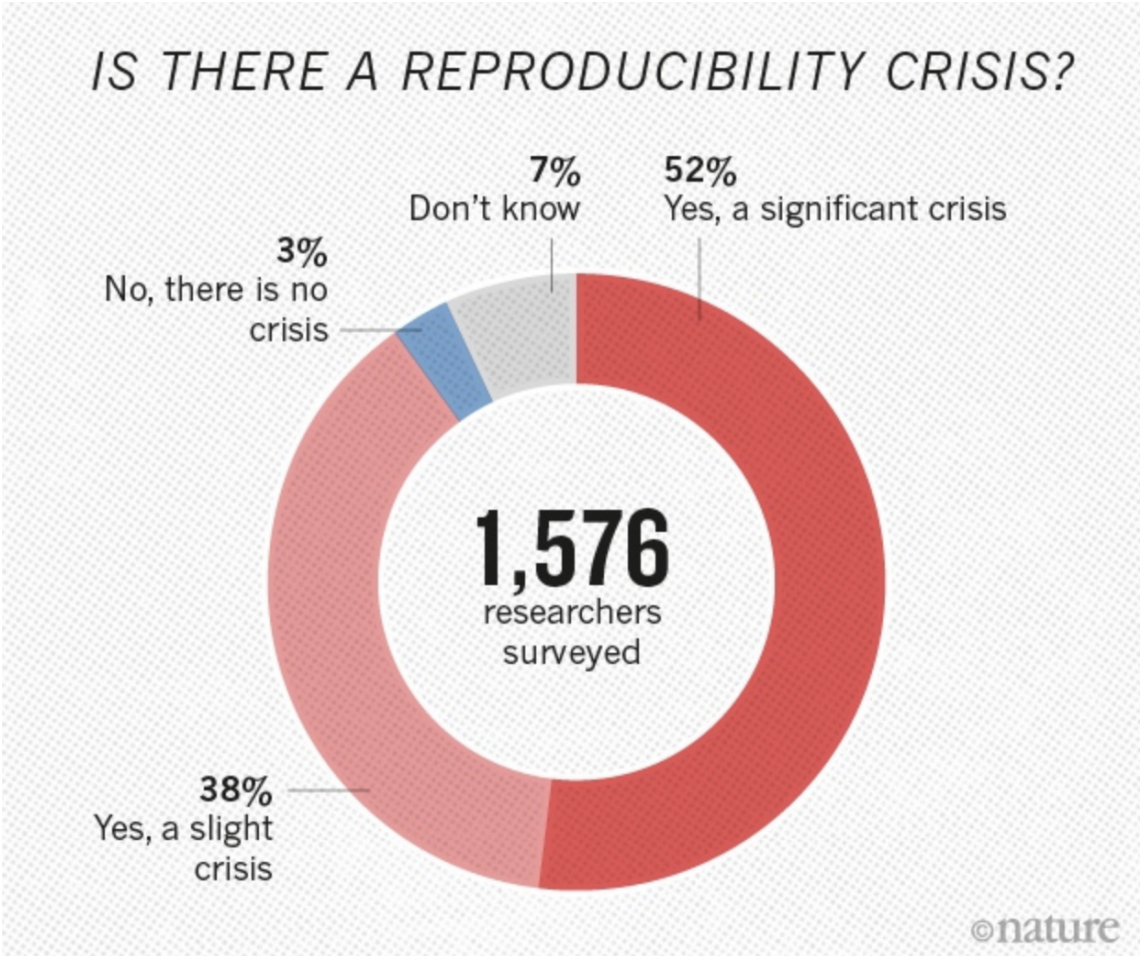
Scientists admit to engaging in questionable research practices.



Fanelli *PLoS ONE* (2011)

Scentists think there is a "reproducibility" crisis

or a "slight" crisis? 🤔



Baker *Nature* (2018)

1. Scientific Integrity Problems

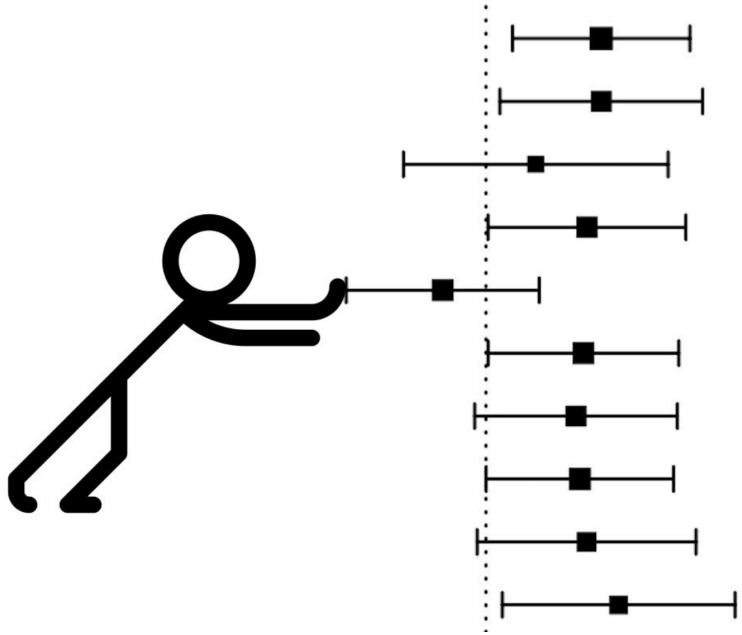
1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

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



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

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.

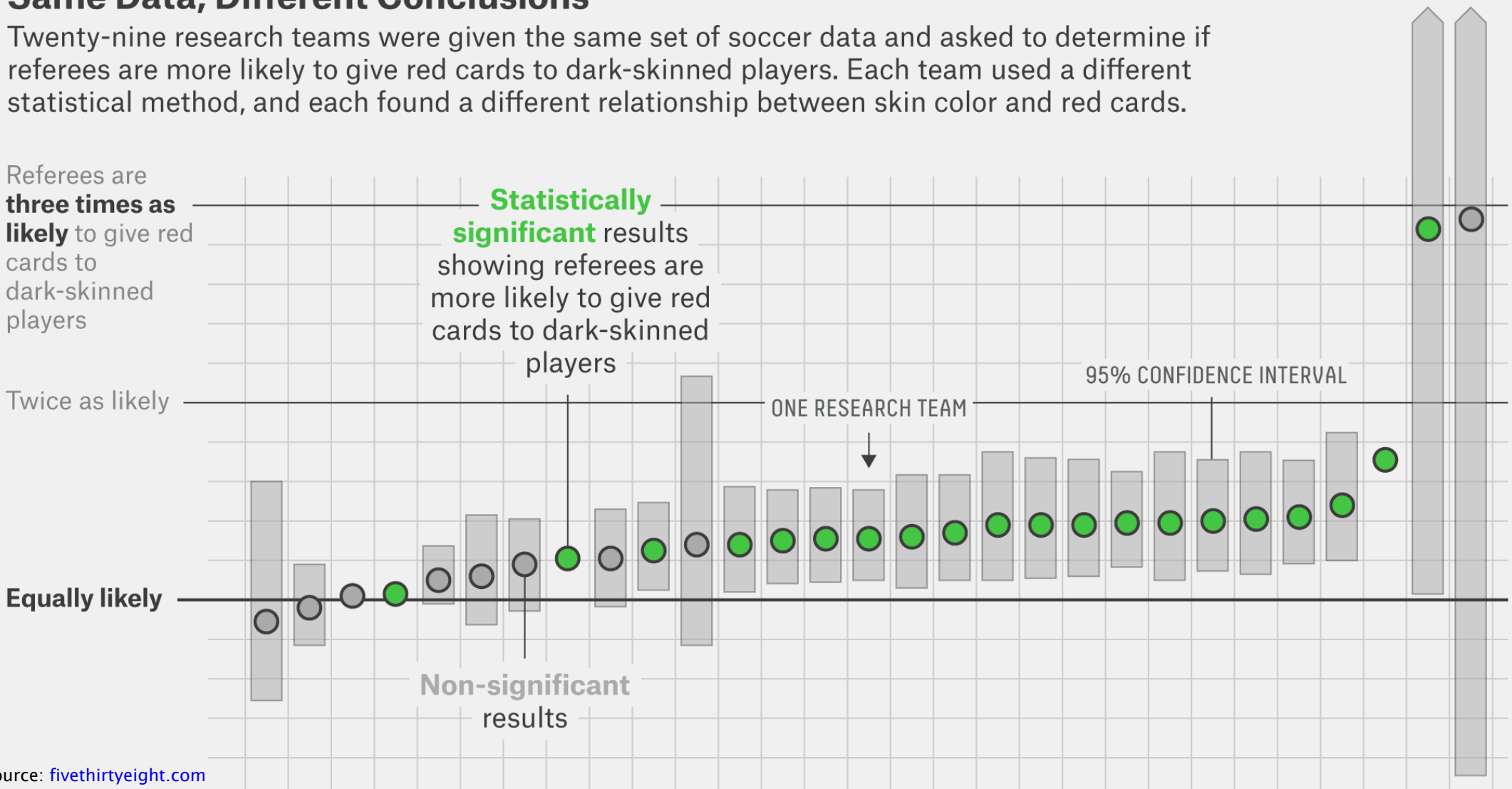
What's the problem?

- 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 next week's election 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?

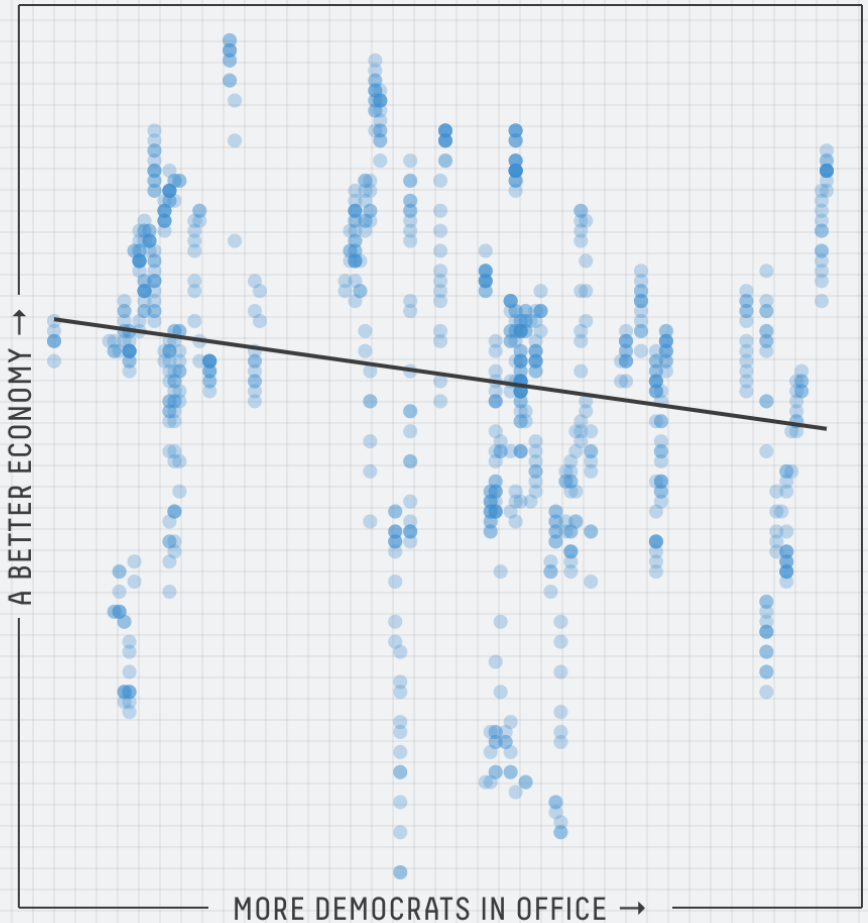
- [x] **Employment**
- [ ] **Inflation**
- [ ] **GDP**
- [ ] **Stock prices**

Other options

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

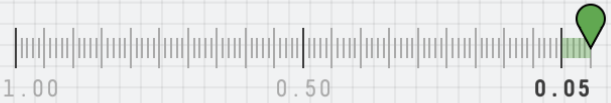
3 IS THERE A RELATIONSHIP?

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?

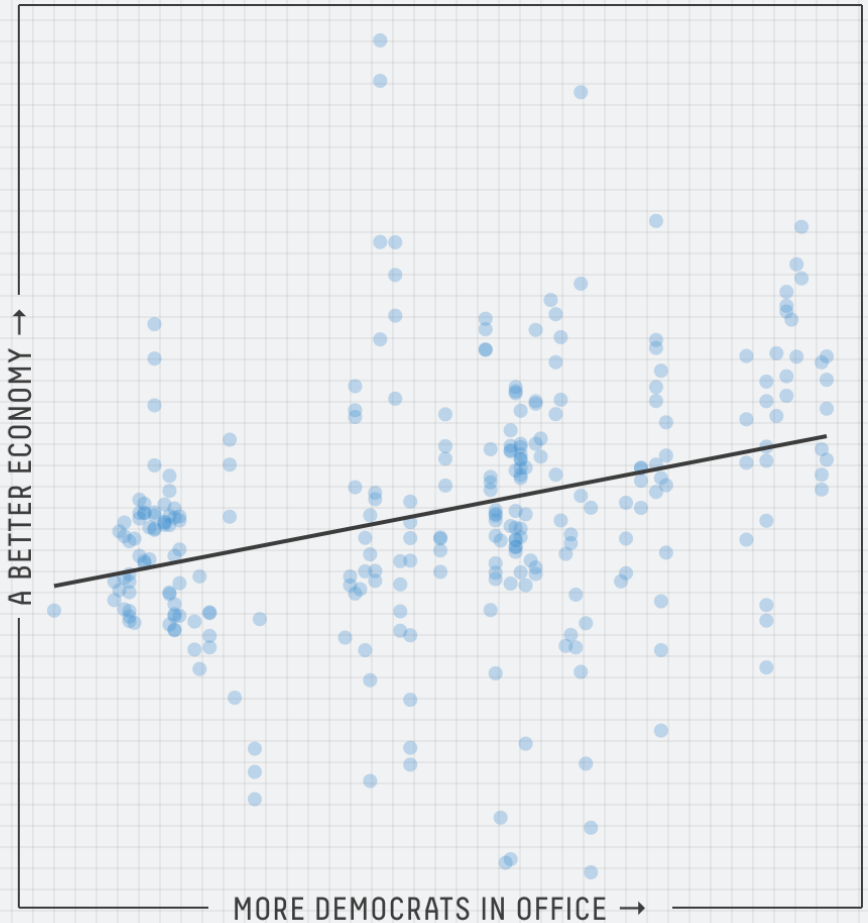
- [ ] **Employment**
- [ ] **Inflation**
- [x] **GDP**
- [ ] **Stock prices**

Other options

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

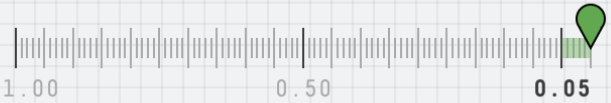
3 IS THERE A RELATIONSHIP?

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

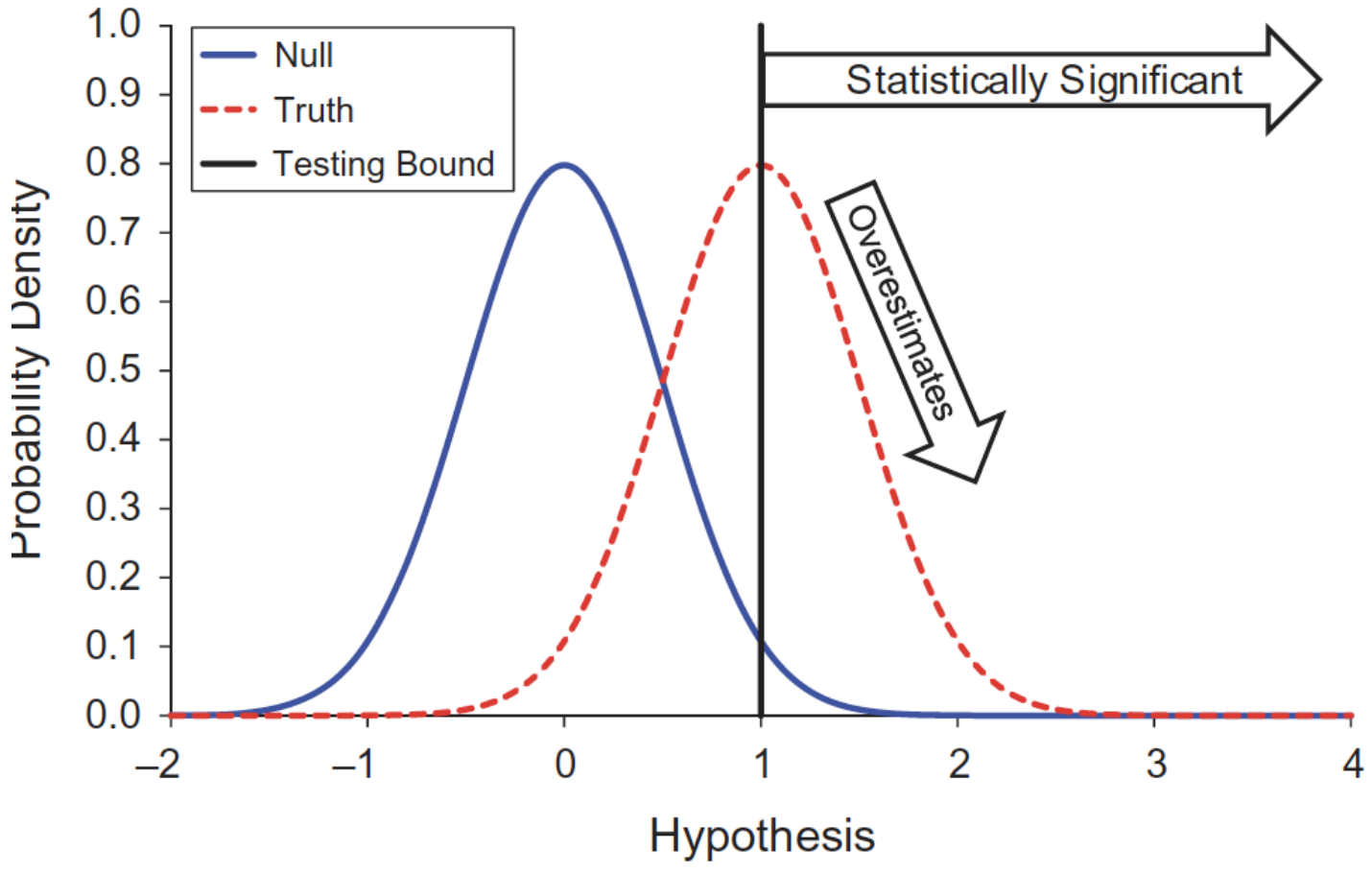
You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **positive** effect on the economy. Get ready to be published!

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 NHST facilitates non-replication

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



Lash (2017)

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.**)

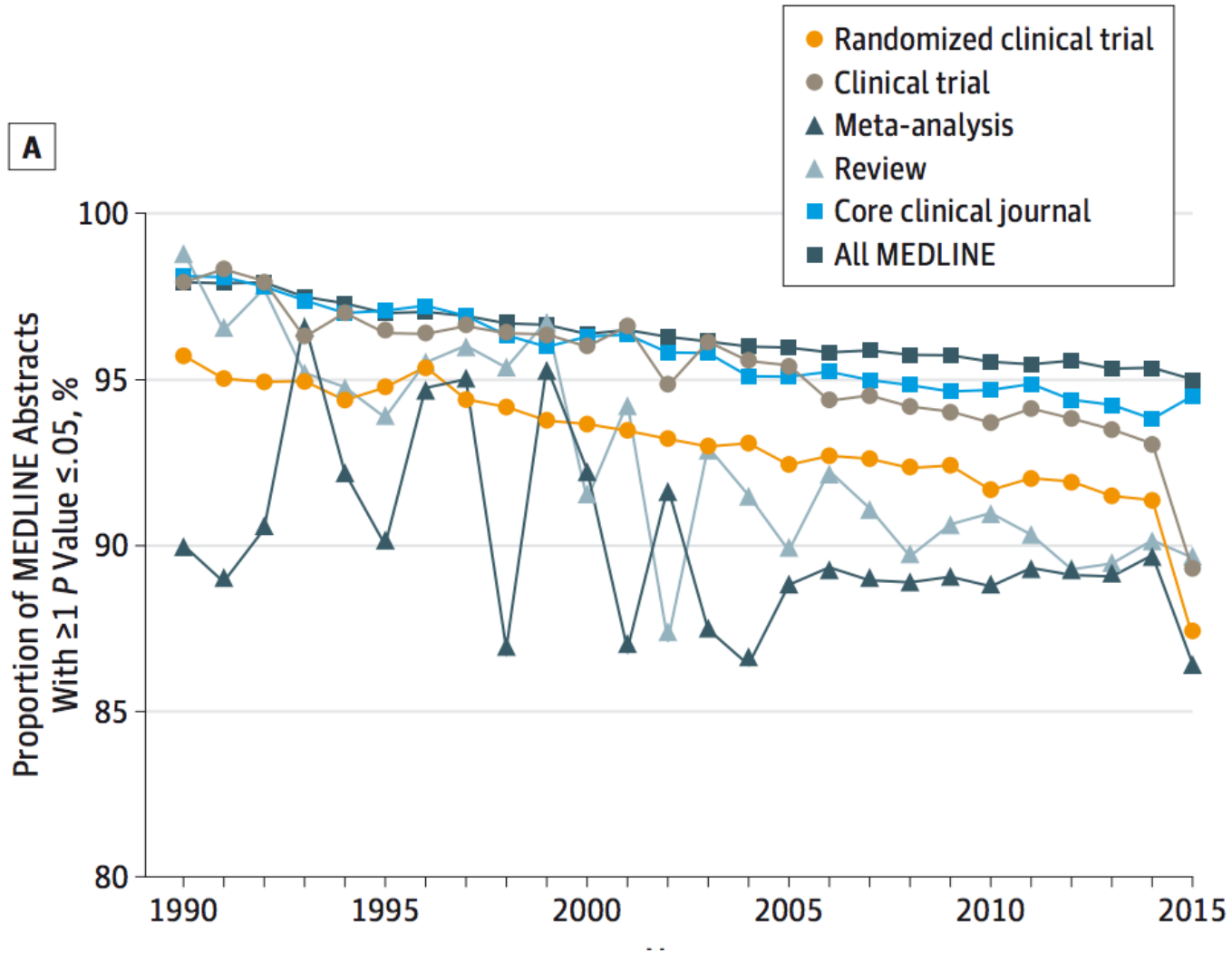
Key Words: atrial fibrillation • testosterone • testosterone replacement therapy

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

How do we know there is p-hacking?

(2) Seriously, everything is significant

P-values in the biomedical literature, 1990-2015



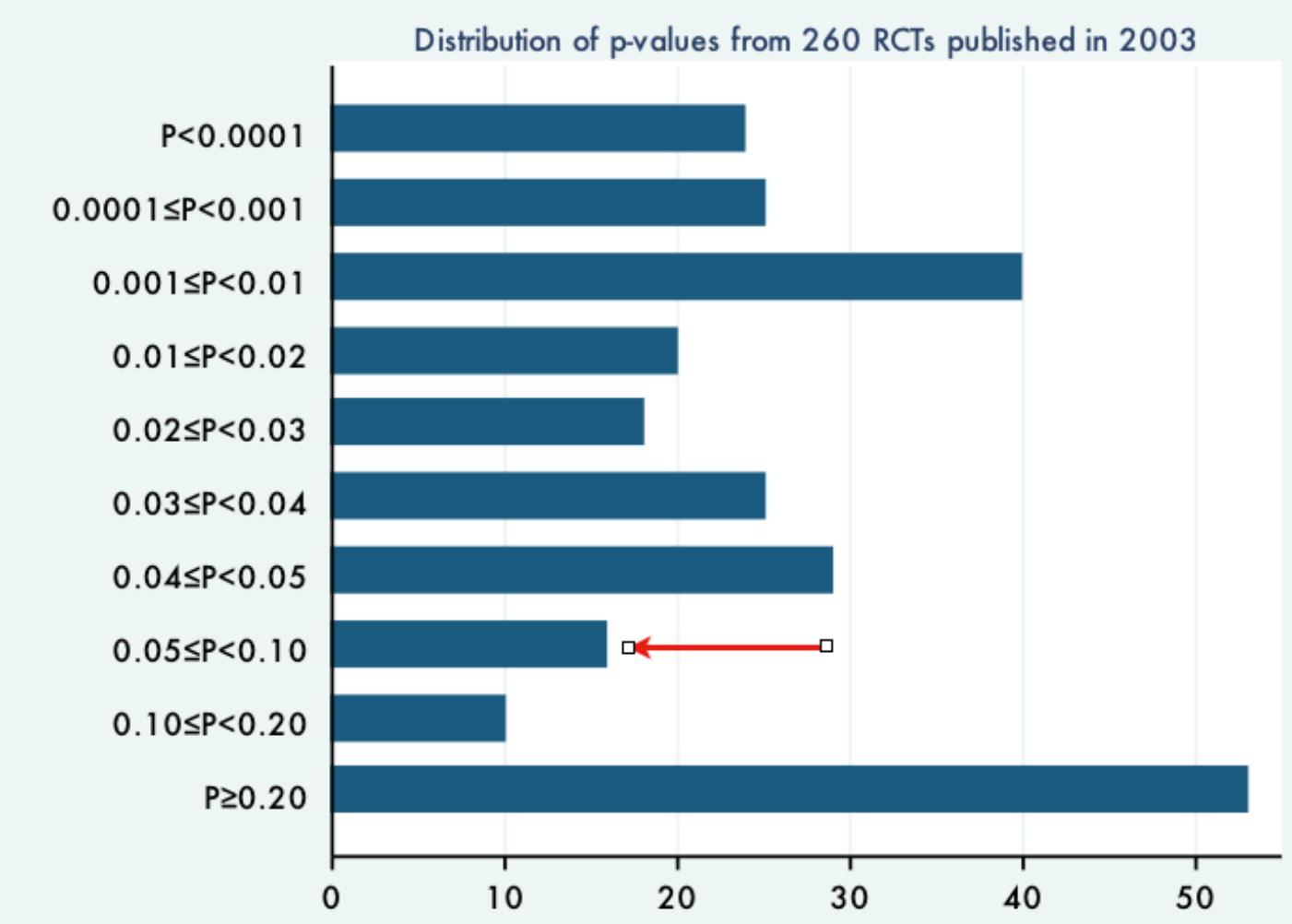
Chavalarais et al. (2013)

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

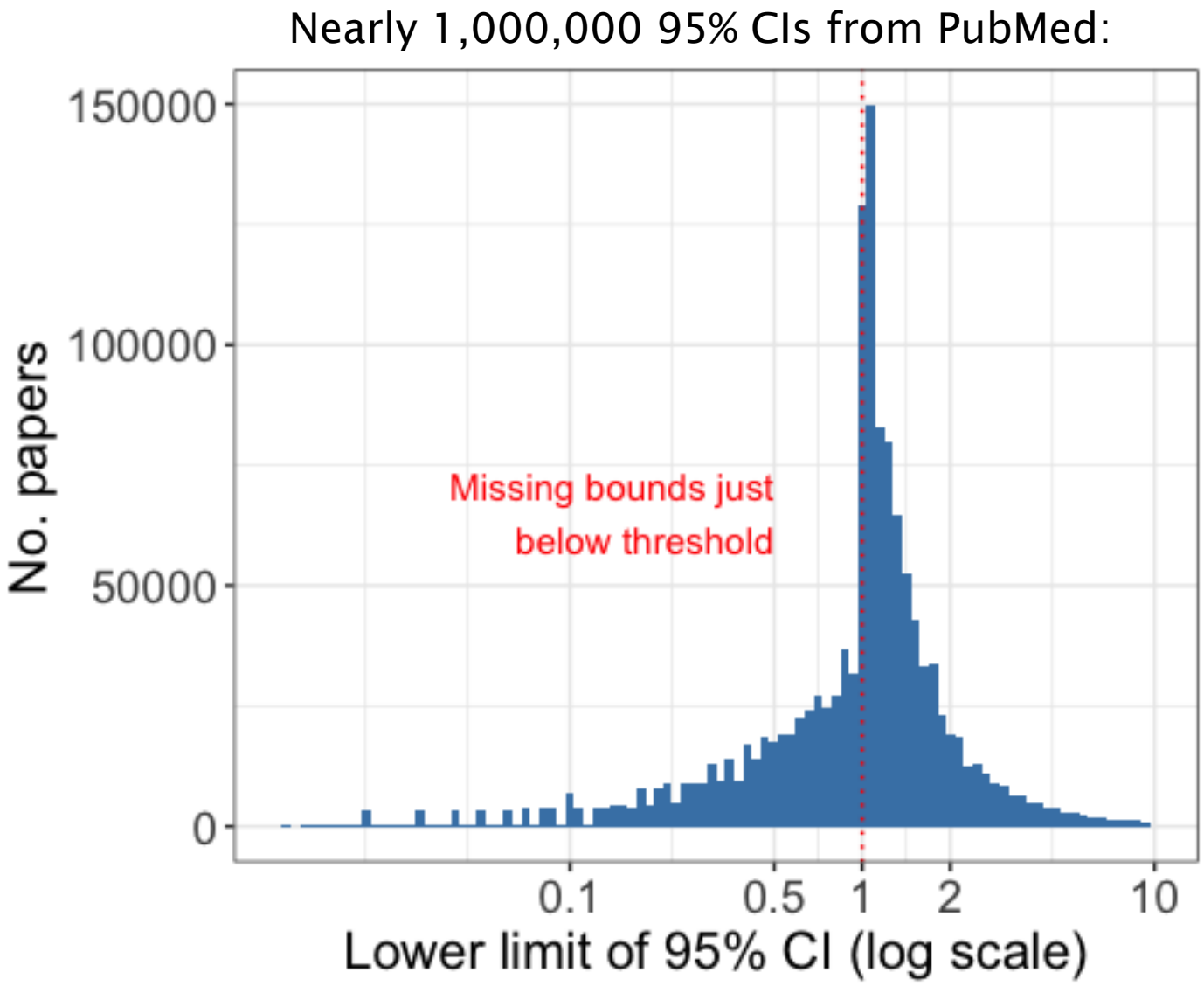


Gotzsche (2006)

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

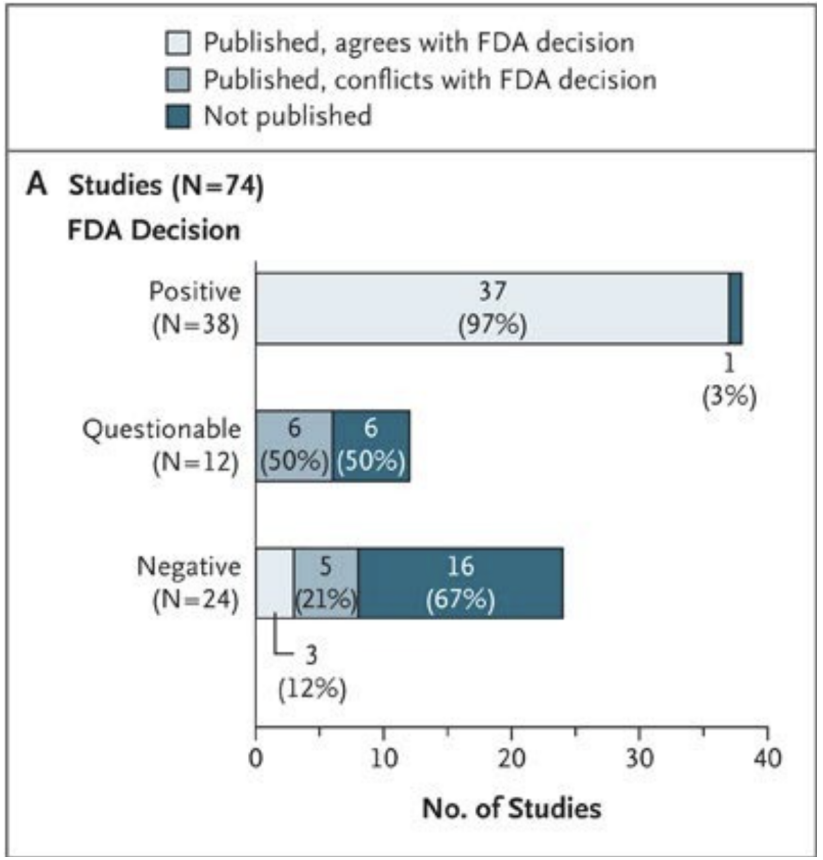
Negative studies of antidepressents less likely to be published.

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)

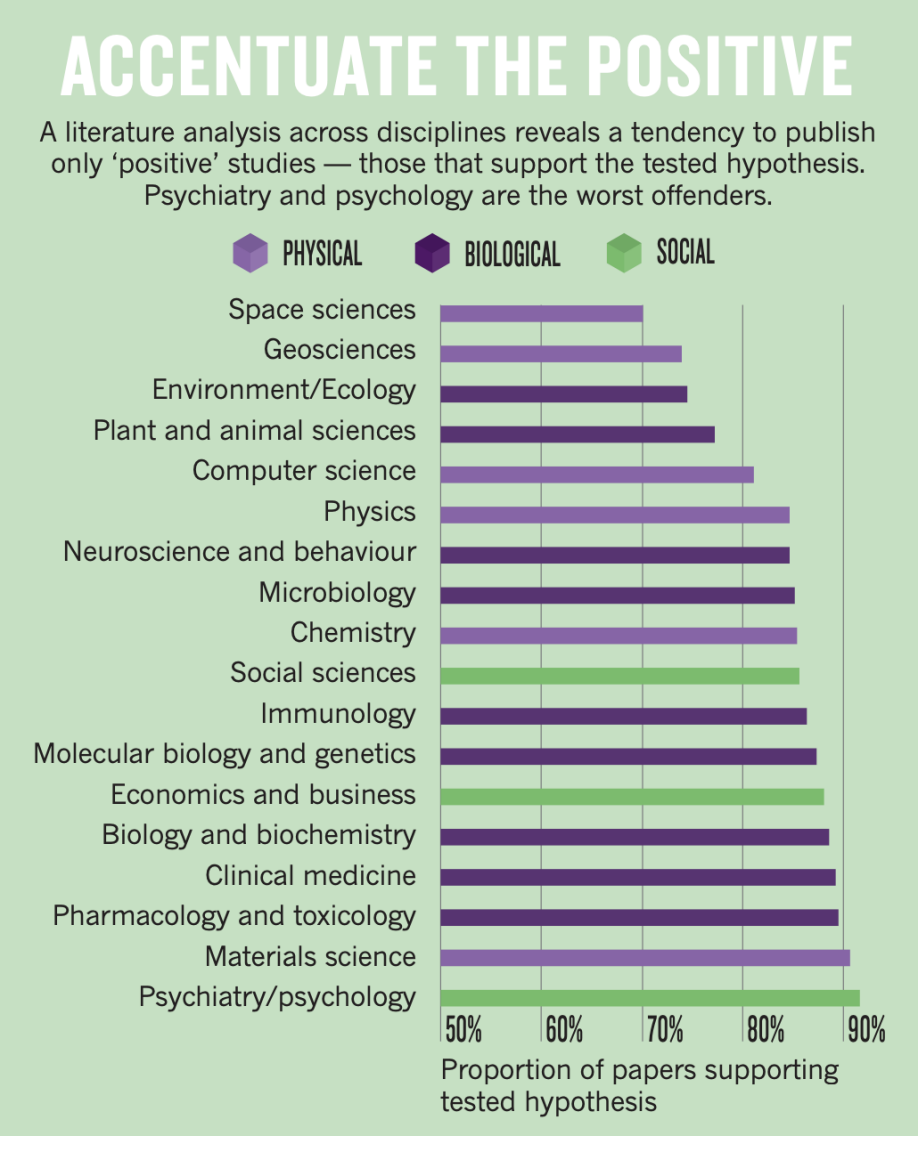
Publication bias affects nearly all disciplines

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.

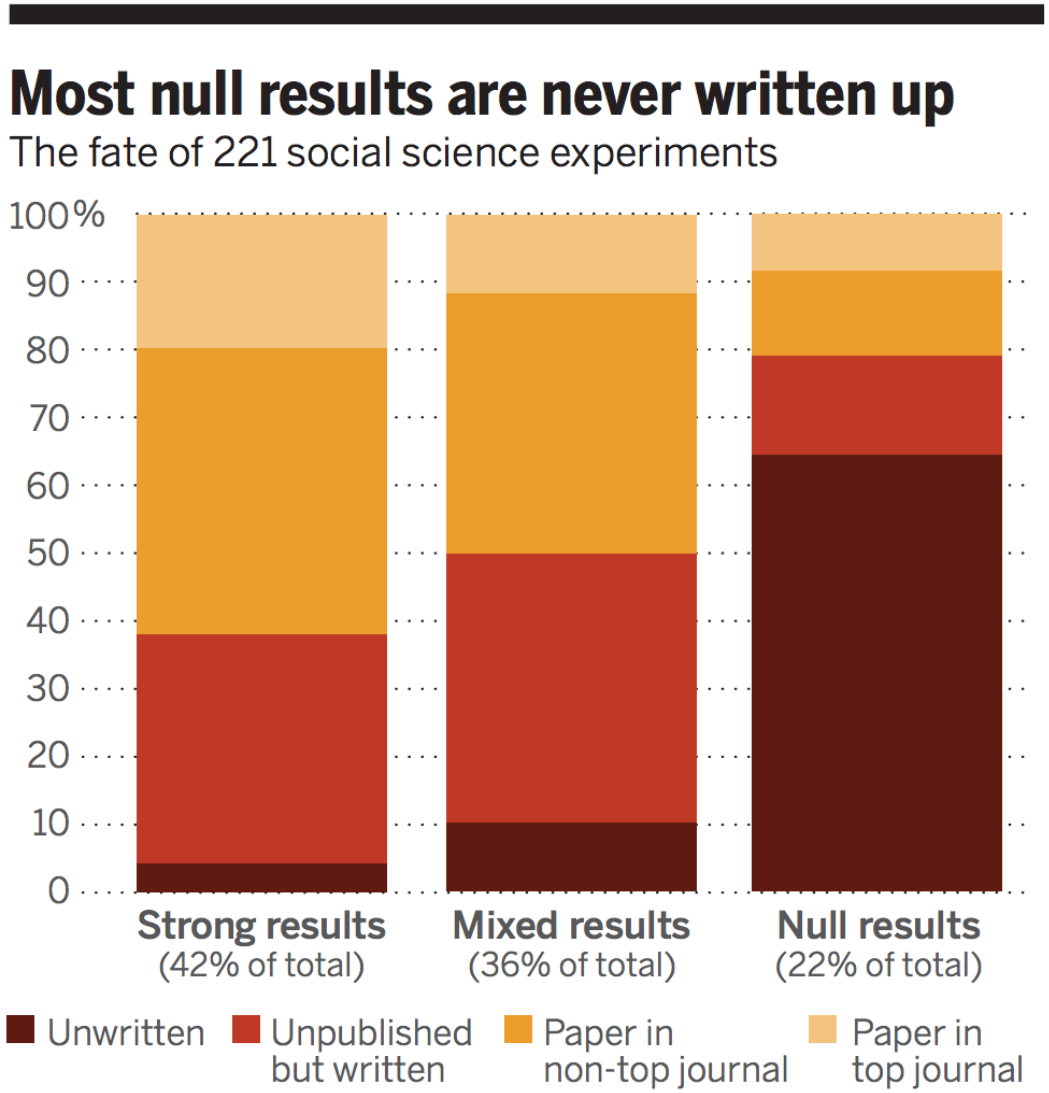


Figure from Mervis in Science 29 Aug 2014;345:992

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

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)

Large scale efforts to replicate studies are not reassuring

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)

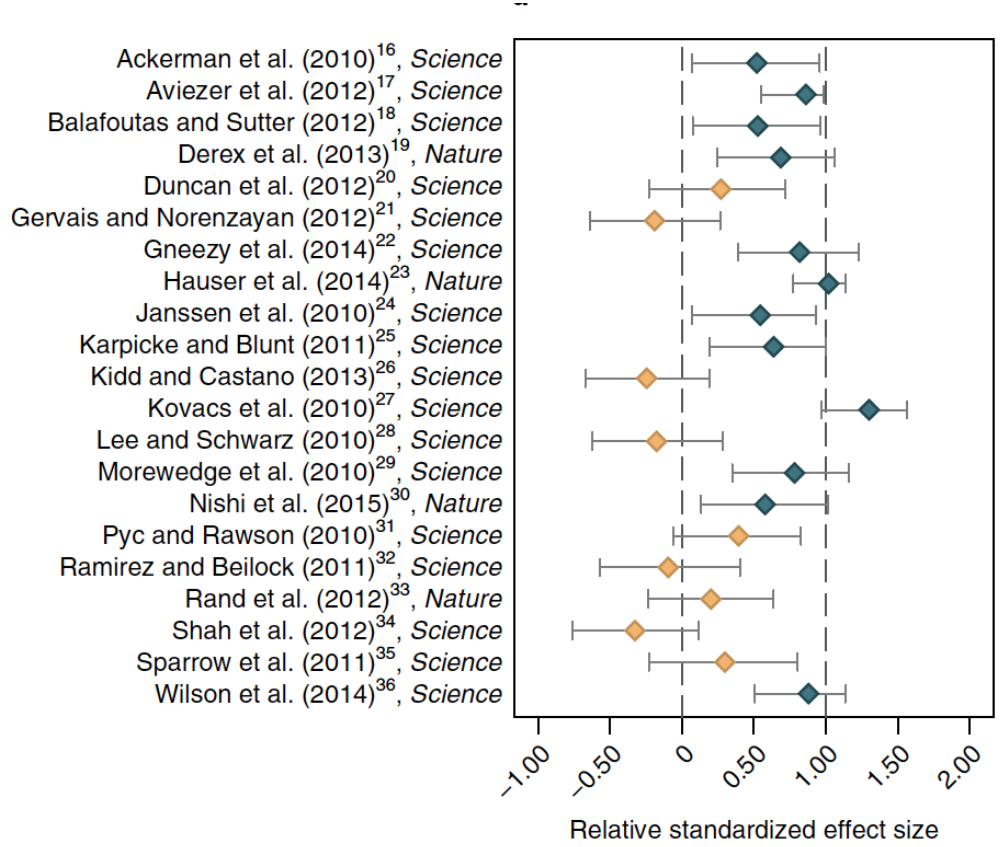
Effect sizes are much lower in replication studies.

Nosek et al. (2017)

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 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.

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

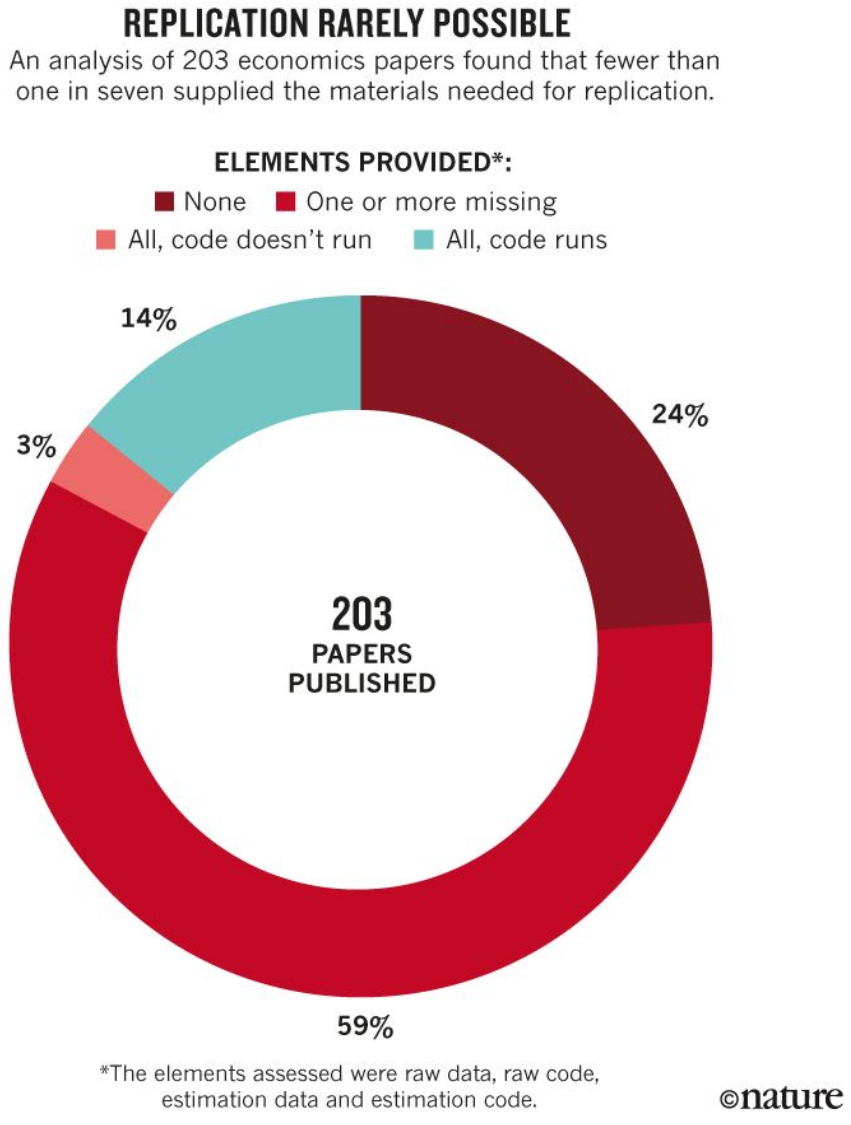
Miyakawa *Molecular Brain* (2020) 13:24

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

1. Scientific Integrity Problems

1.1 Mertonian norms

1.2 Significance testing

1.3 Non-replication

1.4 Incentive structure

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

Incentive problems

Remember Brian Wansink?

After encouraging his postdoc to "find" specific results, fish for interactions, change the dependent variable, and eliminate outliers, he concluded:

> 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.

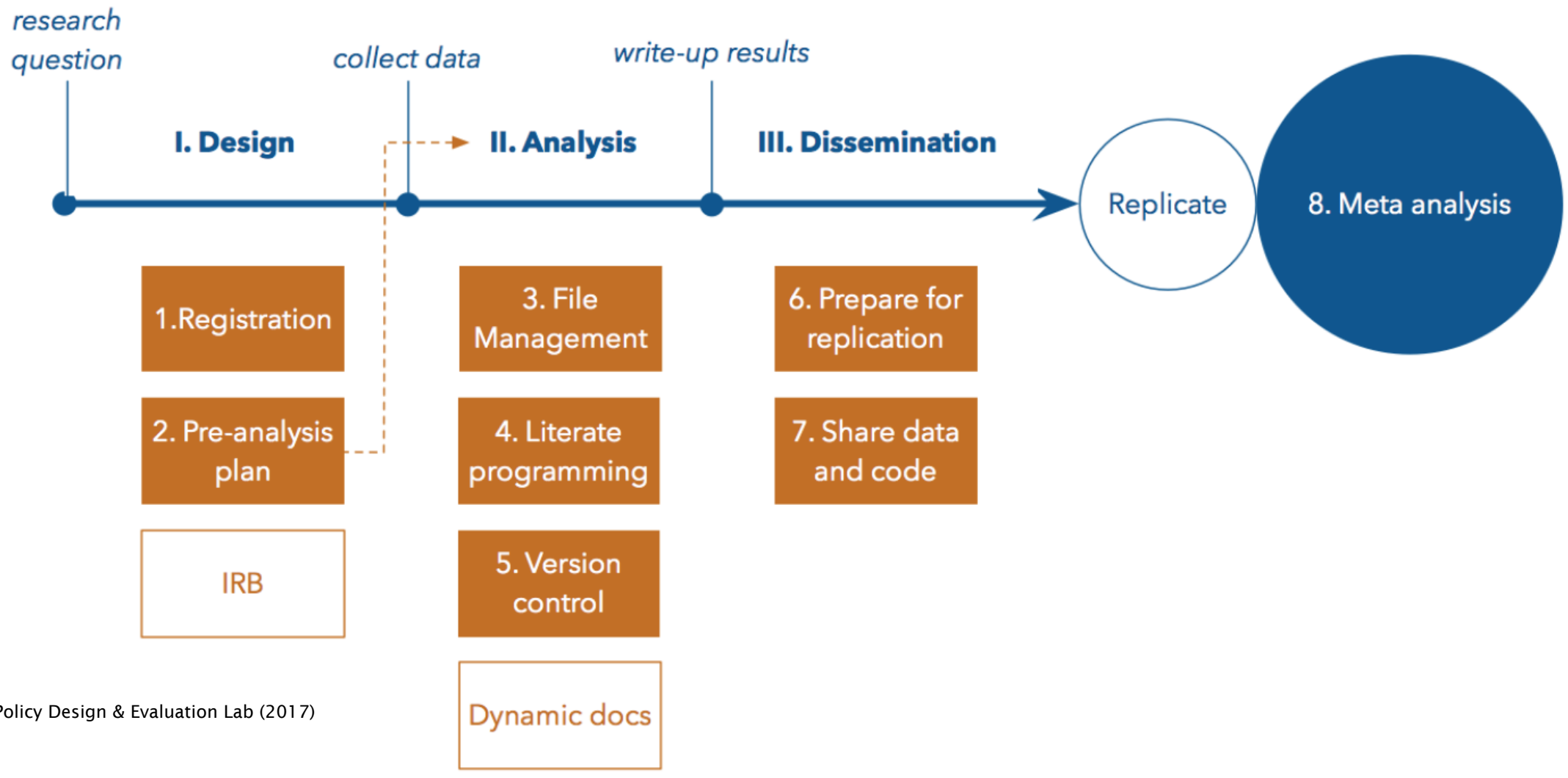
Summary points

Science is conducted by humans.

Many counternorms exist that undermine scientific integrity.

What can we do about it?

A reproducible path forward: Reminaging the research lifecycle?



Policy Design & Evaluation Lab (2017)

Break! ☕

10:00

Reproducible Research: Why and How

Part 2: Design Solutions

Sam Harper

SER Pre-Conference Workshop
2020-10-30

2. Design Solutions

2.1 Preregistration

2.2 Pre-analysis plans

2.3 Reporting guidelines

2. Design Solutions

2.1 Preregistration

2.2 Pre-analysis plans

2.3 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 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 not preregistration?

- Observational studies are hard.
- Manuscripts may adhere to registrations rather than reality.
- May discourage innovation/exploration.
- Pre-specification is irrelevant to the credibility of inference.
- Severe tests of hypotheses are more important than pre-specification.

COMMENTARY

Should Preregistration of Epidemiologic Study Protocols Become Compulsory?

Reflections and a Counterproposal

Timothy L. Lash[a,b] *and Jan P. Vandenbroucke*[c,d]

There is an ongoing debate regarding preregistration of epidemiologic study protocols.[1–4] We examine the basic idea that preregistration of study protocols and their associated hypotheses would enhance the reliability of observational research. We define instances in which preregistration would be useful, and we support a counter-proposal: a public registry containing descriptions of collected epidemiologic data.

A decision to institute compulsory preregistration of protocols for observational studies—to be enforced by editors and reviewers as sometimes suggested[1–3]—is not to be taken lightly, and should not be endorsed solely on the basis of an analogous system instituted for randomized trials. Negative reactions toward compulsory registration have been published elsewhere.[5–12] Note that it is the compulsory preregistration of protocols that is most at issue. There are already mechanisms by which epidemiologists can voluntarily preregister their protocols,[13] if they feel preregistration is advantageous. The open question is whether such preregistration should be required in order for observational research to be published in leading journals (assuming the same "enforcement" mechanism would be adopted as for clinical trials).

We examine the validity of the analogy between randomized trials and observational studies with regard to the value of preregistering protocols. We then examine the idea that prespecification of a hypothesis enhances the credibility of results, and that avoidance of "false positives" should always be a primary concern. We discuss research settings when preregistration of an observational study protocol might be of value. Finally, as a

Lash and Vandenbroucke, *Epidemiology* (2012)

Where can you pre-register your study?

- The AEA registry includes the option to upload PAPs. Search under "advanced options" for studies which include PAPs.
- EGAP is a registry for political science studies, some of which include pre-analysis plans.
- 3ie also has a database (RIDIE) of ongoing international development impact evaluations.
- The Open Science Framework (OSF) also invites pre-registered studies and PAPs.
- Many clinical trials in the US must be registered with Clinicaltrials.gov.

Where can you find templates for preregistration?

See the page at Open Science Foundation

Writing up pre-registered studies

1. Include a link to the registration

2. Report *all* pre-registered results.

3. Explain and justify deviations.

4. Non-registered analyses appropriately described as "exploratory" or "hypothesis generating".

Why does preregistration matter?

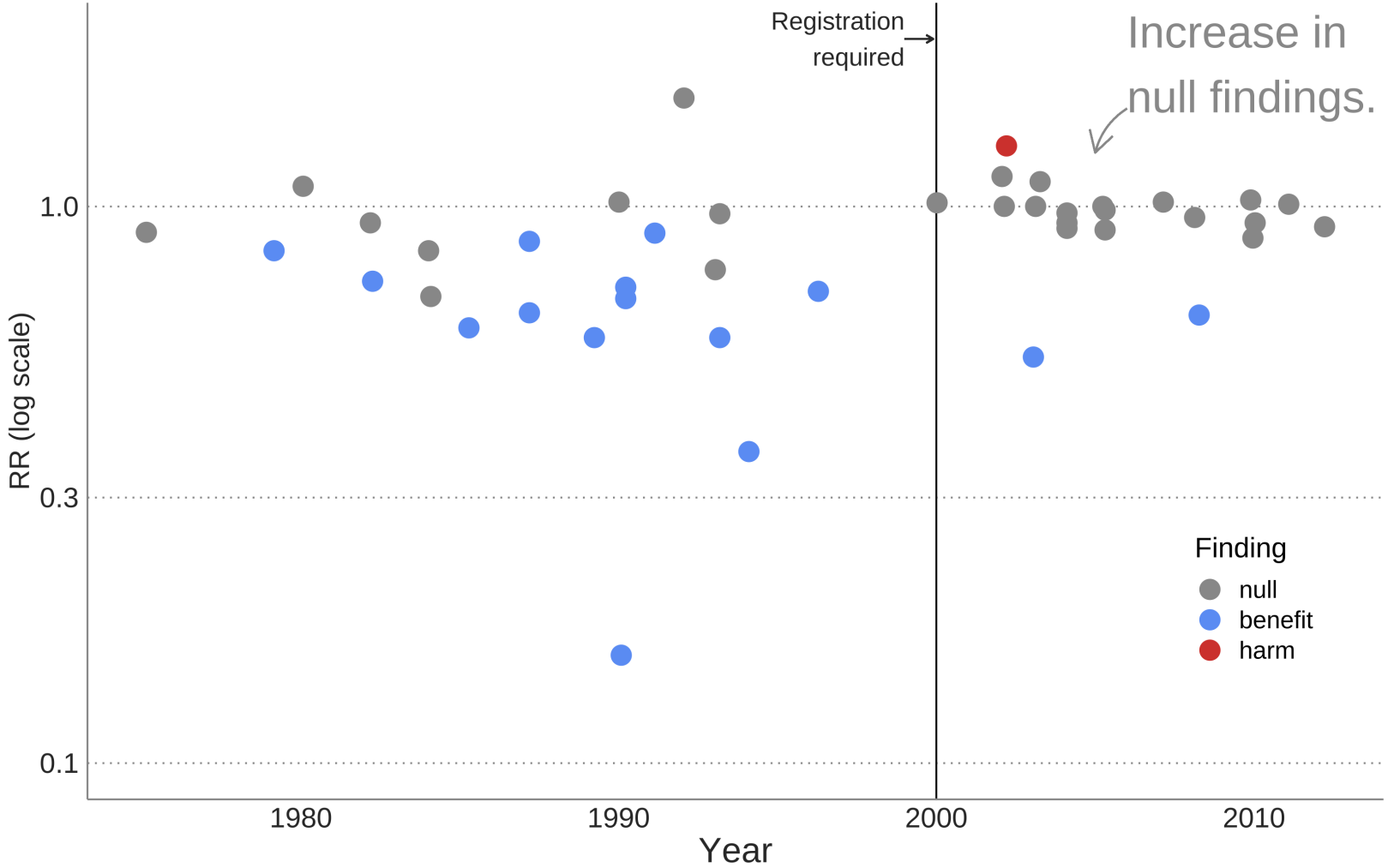
Evidence synthesis should be on *all* the evidence.

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.



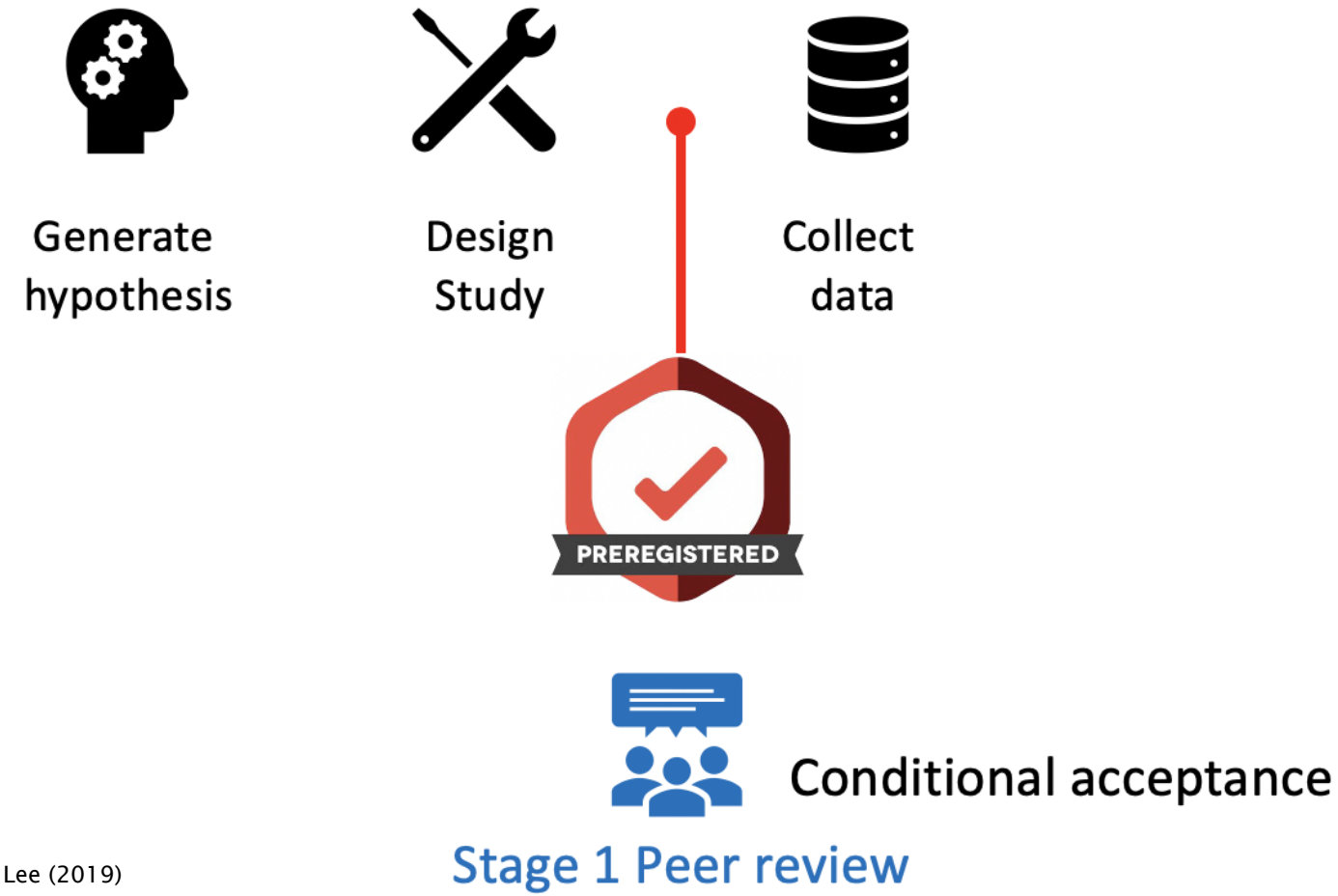
redrawn from Kaplan and Irwin (2015)

What if my results are null?

You showed us that they won't get published!

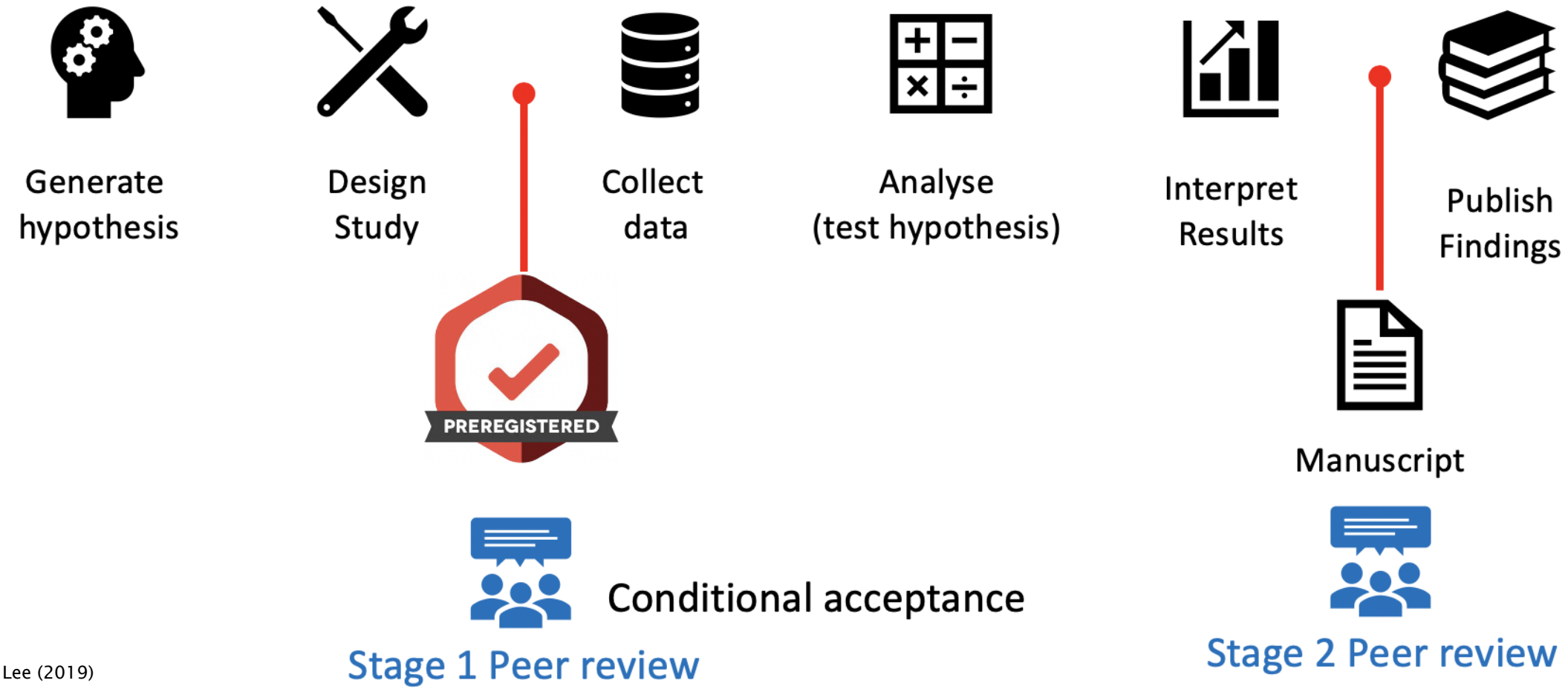
I have to make rent, you know.

Emphasis on design: Registered Reports



Lee (2019)

Emphasis on design: Registered Reports

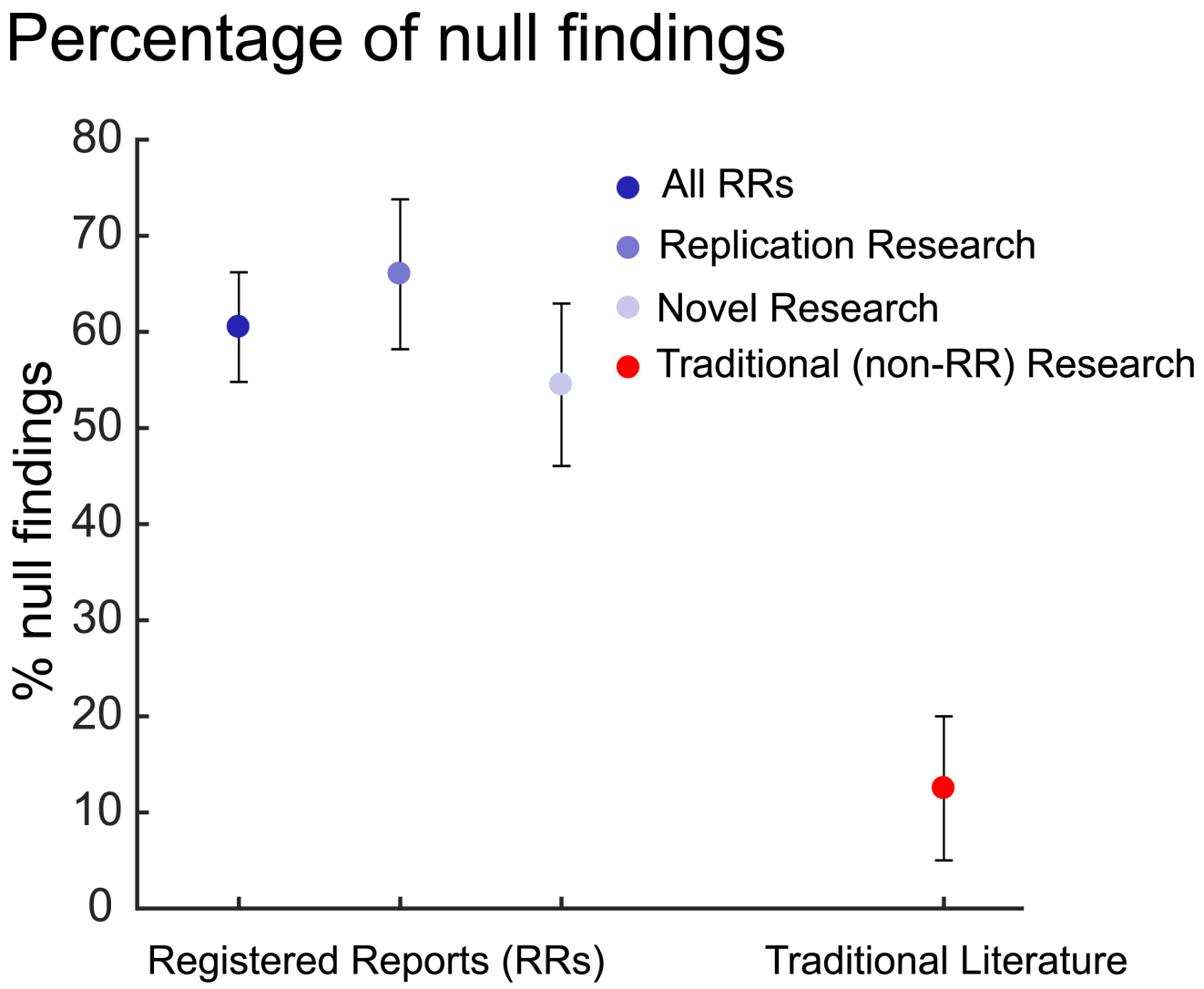


Lee (2019)

RRs in Psychology

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

Big difference from non-registered studies.

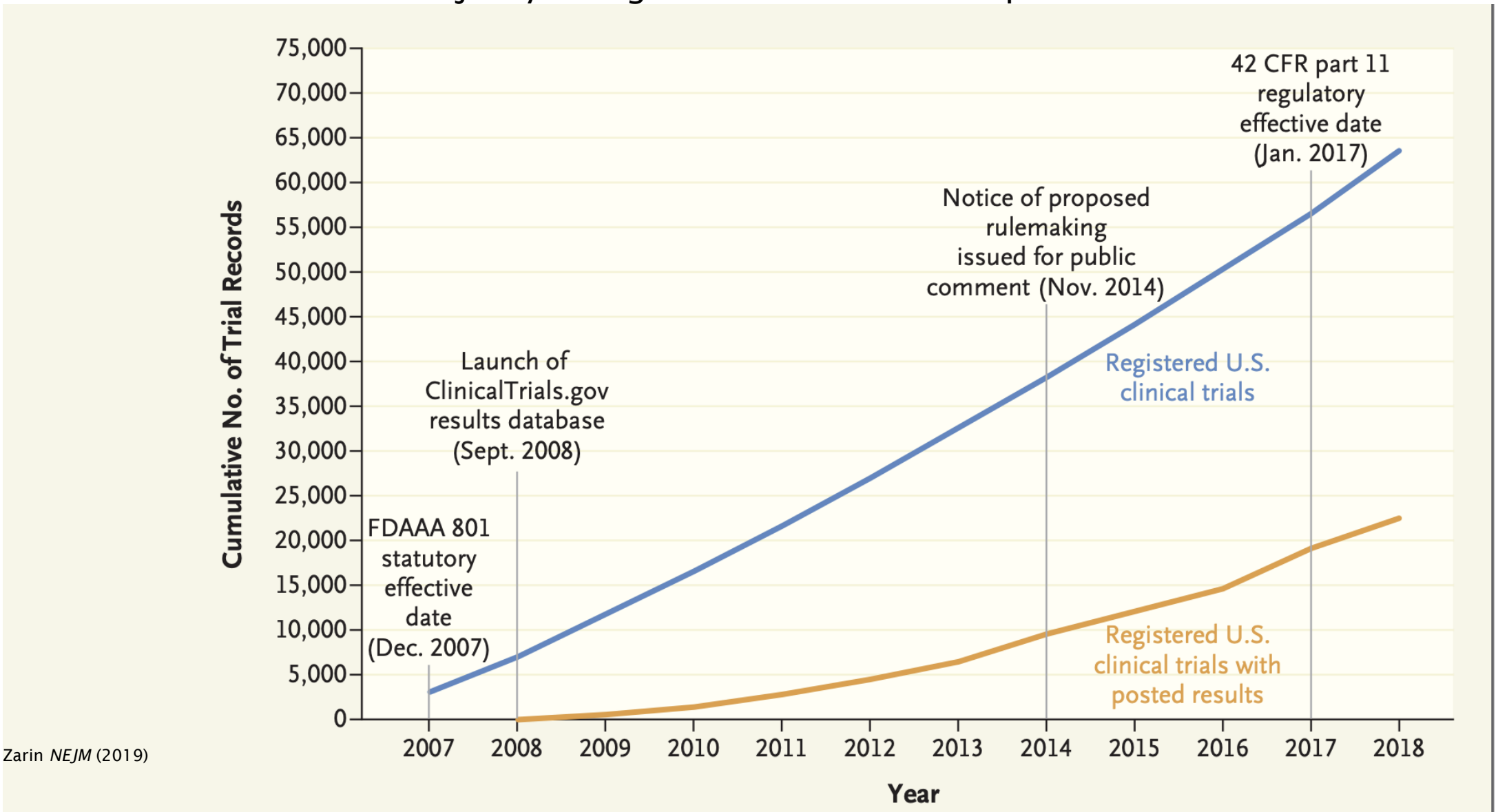


Allen & Mehler, *PLoS Biology* (2019)

Registration is useful

but not sufficient

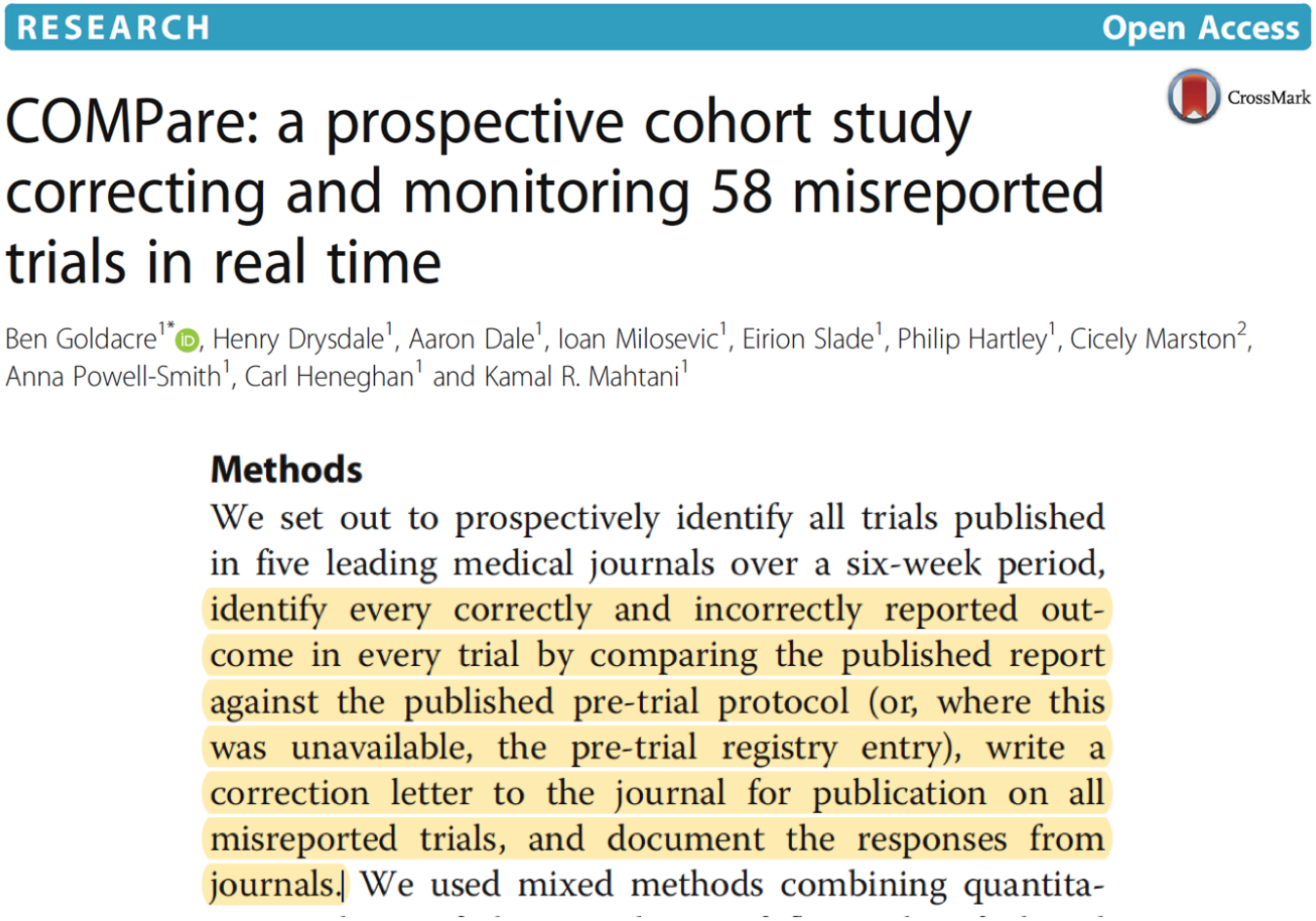
A majority of registerd RCTs still not reported.



Zarin *NEJM* (2019)

But is preregistration enough?

- Still many differences between registration and published reports.



RESEARCH Open Access

CrossMark

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)

2. Design Solutions

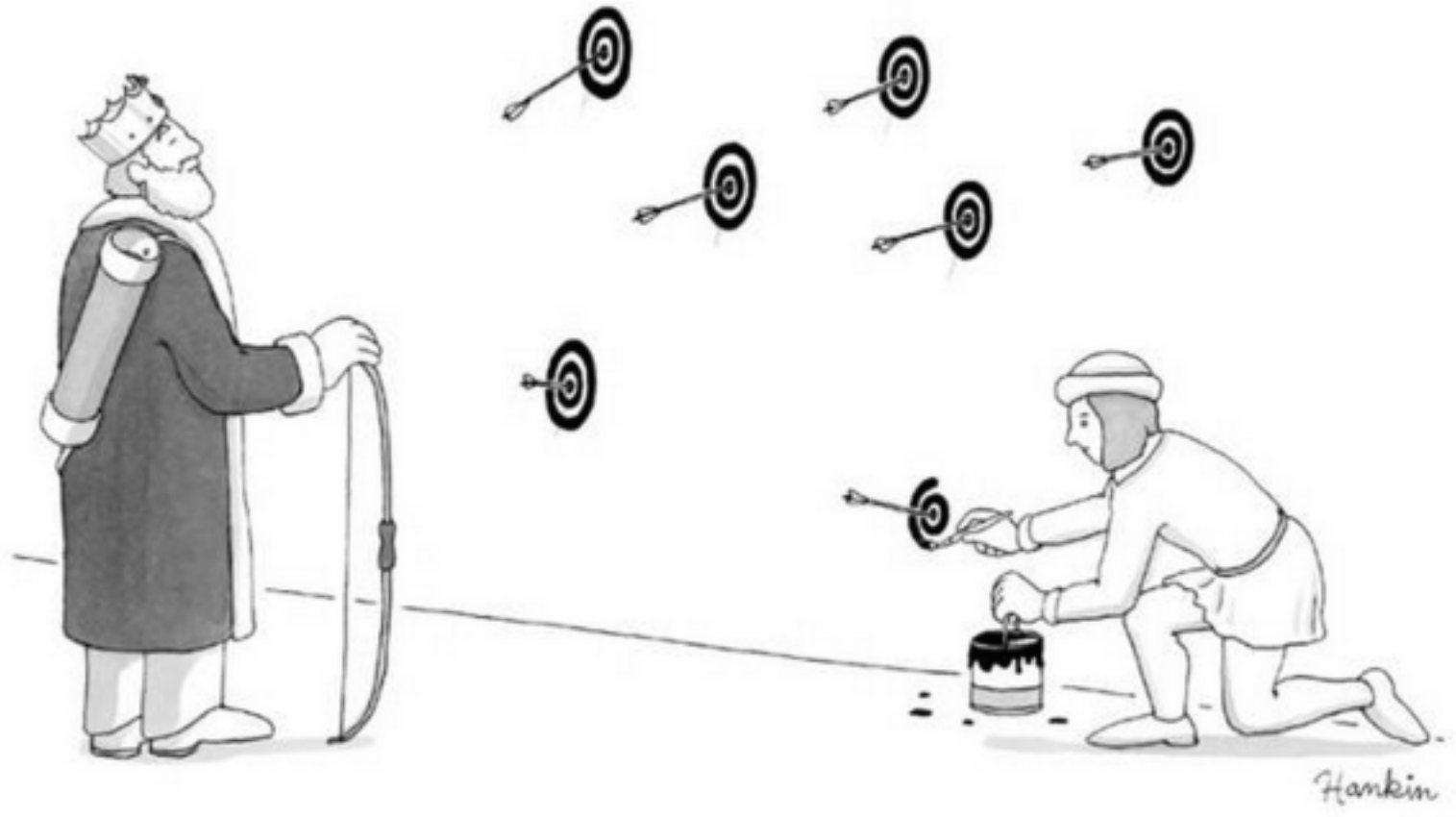
2.1 Preregistration

2.2 Pre-analysis plans

2.3 Reporting guidelines

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 development economics

RESHAPING INSTITUTIONS: EVIDENCE ON AID IMPACTS USING A PREANALYSIS PLAN*

KATHERINE CASEY
RACHEL GLENNERSTER
EDWARD MIGUEL

Despite their importance, there is limited evidence on how institutions can be strengthened. Evaluating the effects of specific reforms is complicated by the lack of exogenous variation in institutions, the difficulty of measuring institutional performance, and the temptation to "cherry pick" estimates from among the large number of indicators required to capture this multifaceted subject. We evaluate one attempt to make local institutions more democratic and egalitarian by imposing participation requirements for marginalized groups (including women) and test for learning-by-doing effects. We exploit the random assignment of a governance program in Sierra Leone, develop innovative real-world outcome measures, and use a preanalysis plan (PAP) to bind our hands against data mining. The intervention studied is a "community-driven development" program, which has become a popular strategy for foreign aid donors. We find positive short-run effects on local public goods and economic outcomes, but no evidence for sustained impacts on collective action, decision making, or the involvement of marginalized groups, suggesting that the intervention

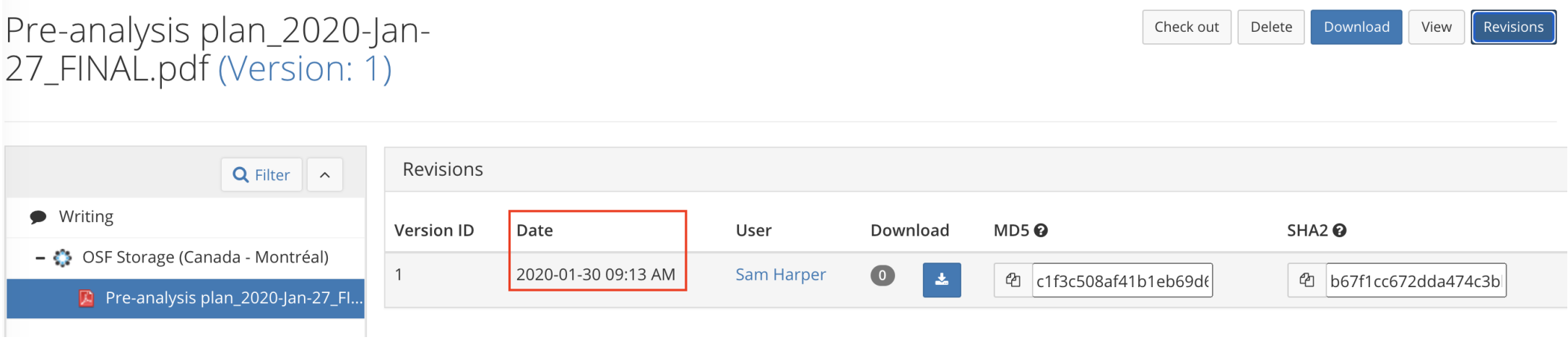
Casey et al. (2012)

Conclusions:

Turning to empirical methods, this paper underscores the importance of PAPs to limit data mining and generate appropriately sized statistical tests, and discusses some of the practical trade-offs we faced in implementation. We confront the fundamental tension between researcher discretion versus commitment and argue that flexibility to explore questions that arise as the research and project unfold is sometimes desirable yet should only be exercised in tandem with complete transparency over deviations from the ex ante specifications. In the context of a PAP, limited flexibility with full transparency allows the scholarly community to make its own assessments about the credibility of different results. We show how misleading an undisciplined interpretation of treatment effects can be in the absence of a PAP by constructing two opposing and equally erroneous narratives based on our data.

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]

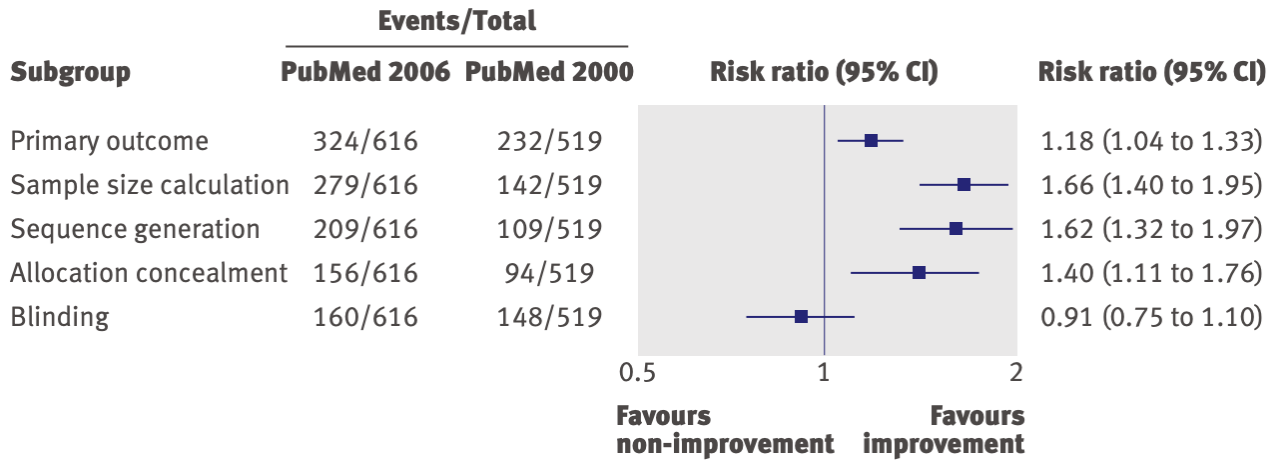


Fig 2 | **Differences in reporting of methodological items between 2000 and 2006**

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

Break! ☕

10:00

Reproducible Research: Why and How

Part 3: Analytic Solutions

Sam Harper

SER Pre-Conference Workshop
2020-10-30

3. Analytic Solutions

3.1 Workflow Management

3.2 Documentation

3.3 Literate Programming

3.4 Version Control

3.4 Dynamic Documents

3. Analytic Solutions

3.1 Workflow Management

3.2 Documentation

3.3 Literate Programming

3.4 Version Control

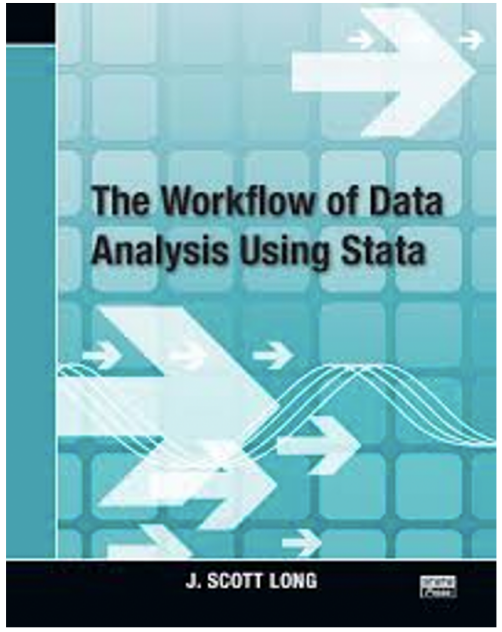
3.4 Dynamic Documents

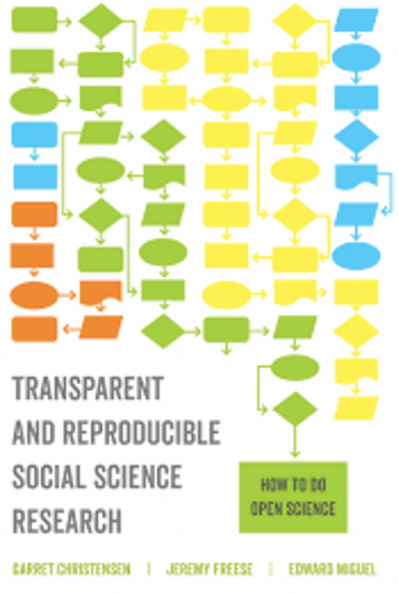
Workflow Advice

Resources for advice on:

- Planning and organization
- Documentation
- Writing / debugging syntax
- Automating your work
- Variable labeling / naming
- Cleaning
- Archiving

Often specific to software but core ideas are the same.



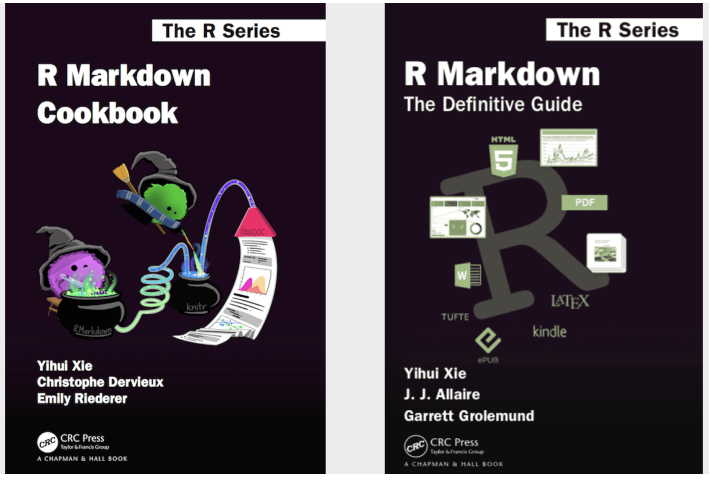


Long (2009); Christensen et al. (2019)

Great online **free** resources for *R*

Technical help for:

- Learning *R*.
- Developing a reproducible workflow.
- Tips & tricks.
- Integration with other software.
- Dealing with frustration.*



*Which is inevitable.

See the series at bookdown.org

Planning your work

Why?

Will save you time.

Plans should help you stay on track.

Hard, and isn't "fun".

What:

- Goals and publishing plans
- Scheduling
- Division of labor
- Datasets needed
- Variable names and labels
- Missing data procedures
- Analysis
- Documentation
- Backing up / archiving

Planning for the entire research pipeline

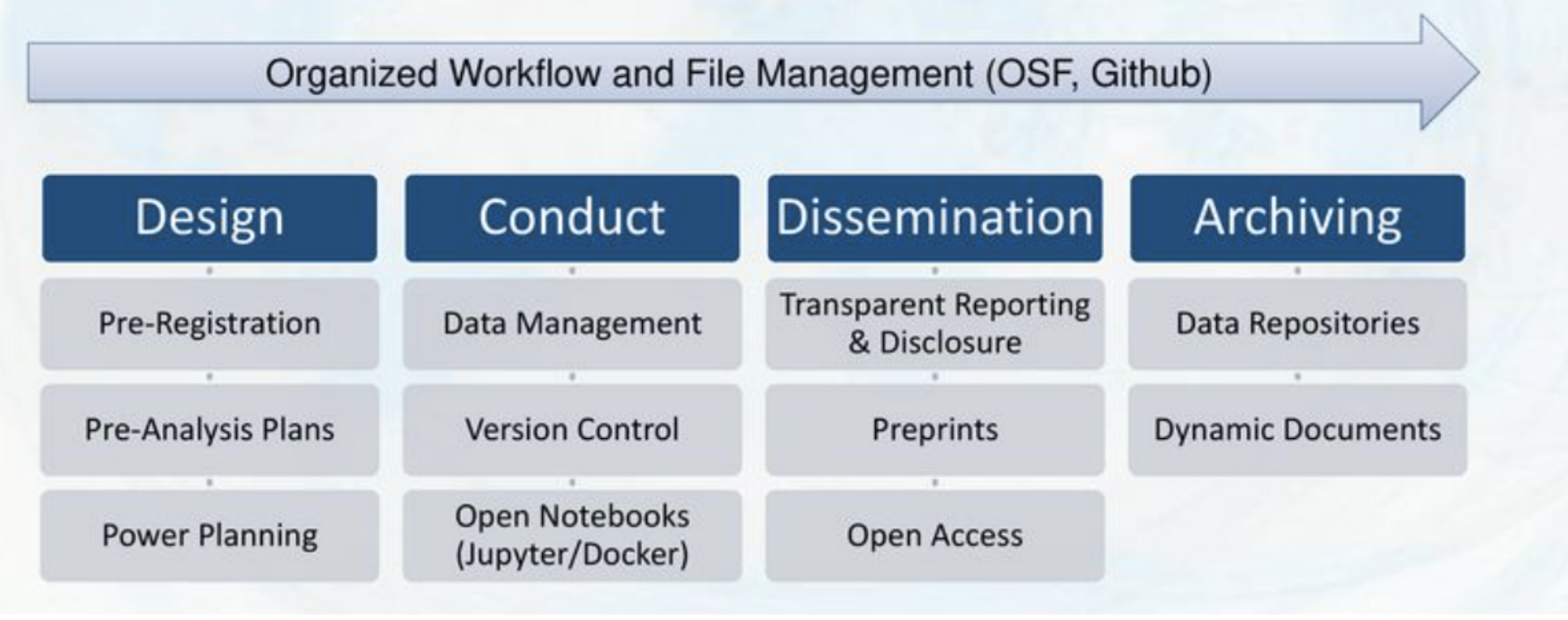
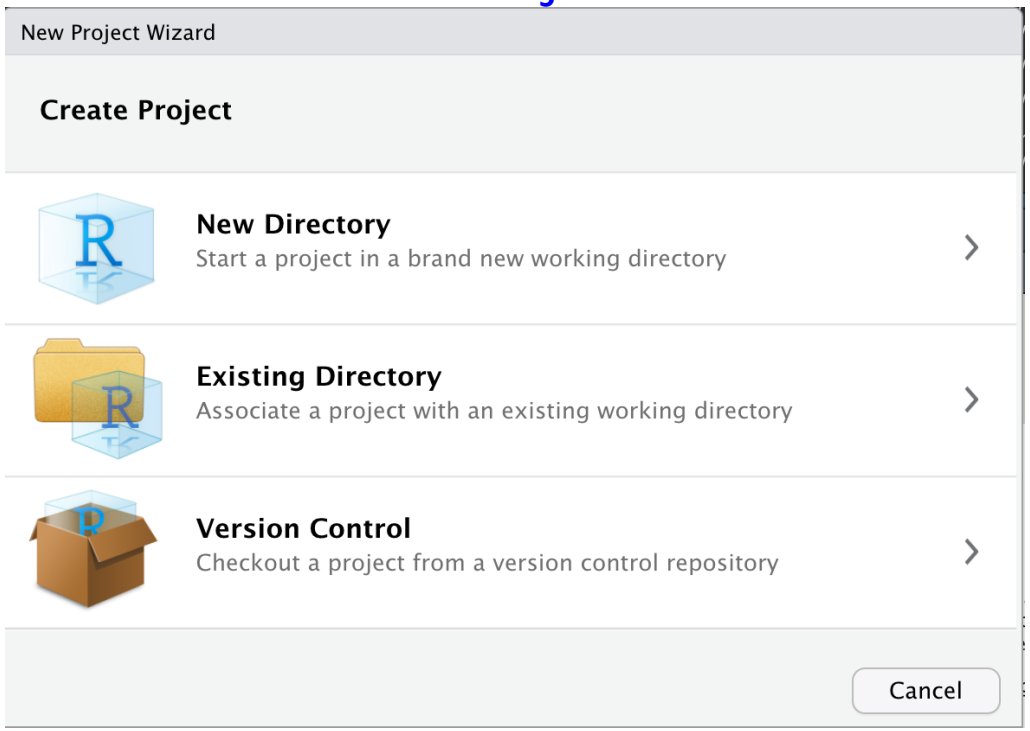


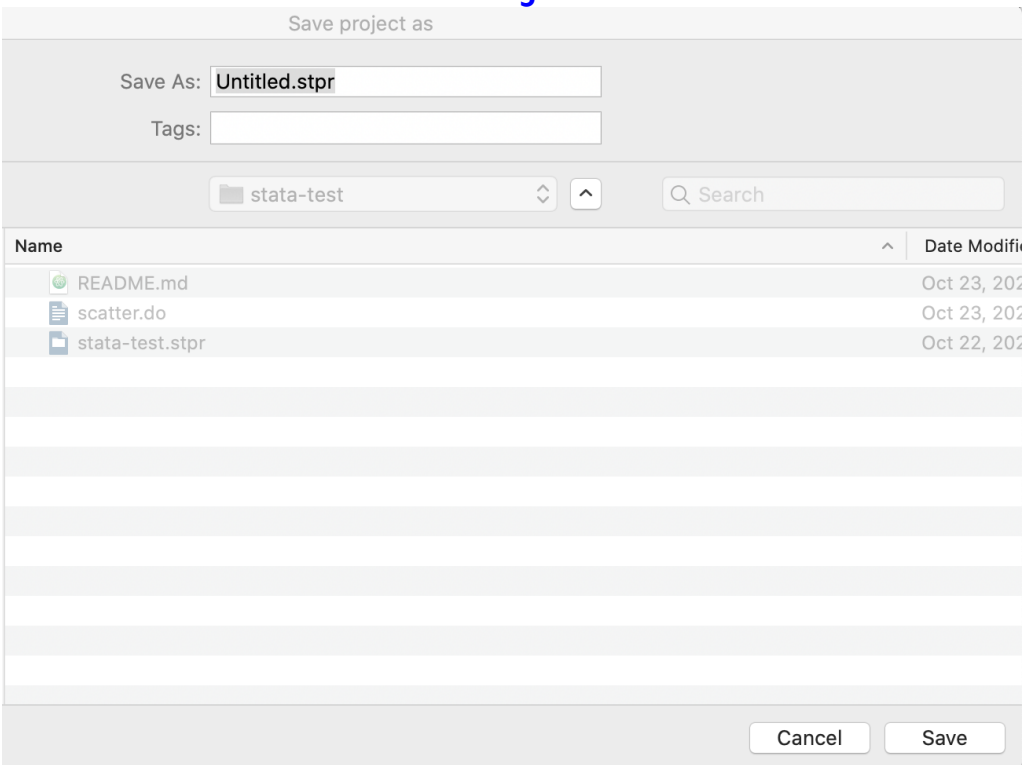
Image credit: OSF

Use projects to keep things organized

RStudio: File > New Project...



Stata: File > New > Project...



Jenny Bryan on Project-oriented workflows. Also possible with SAS.

Why Projects?

By definition, the directory `/User/samharper/project/data` is not reproducible.
But `project/data` works anywhere.

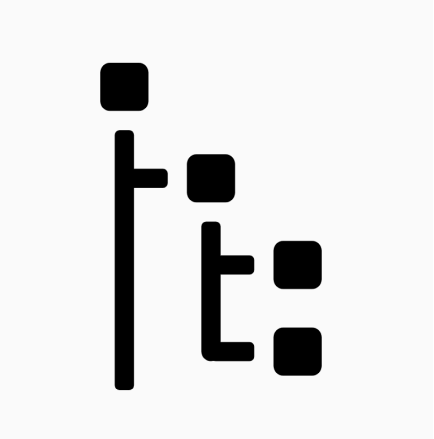
Self contained and portable

Easy to integrate with version control

Working directory set at project location when opened

Fresh session when project is launched

File organization



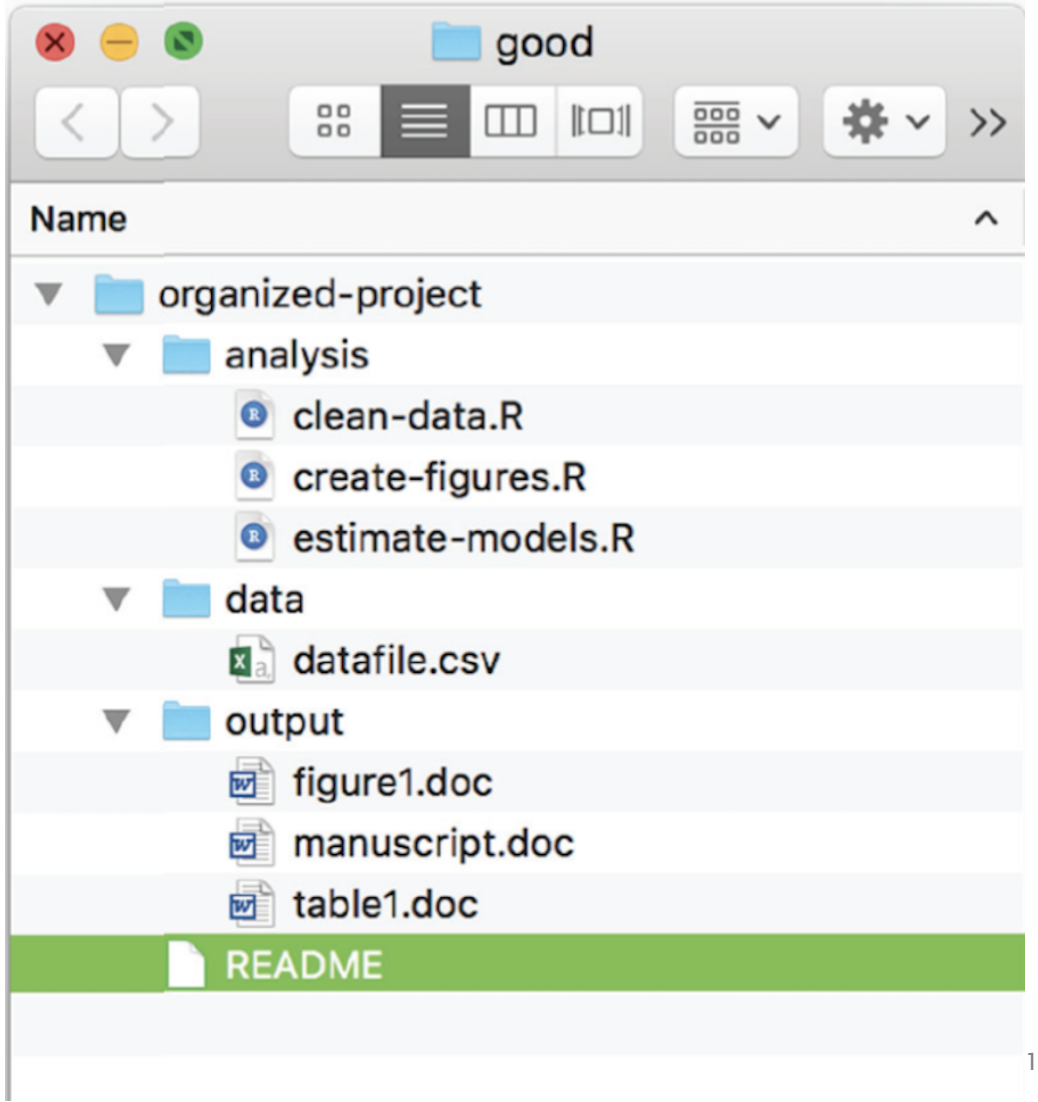
Core idea is to separate raw data, analytic data, code, and outputs.

Why?

- Raw data is **never** altered.
- Separating code from data streamlines re-analysis.
- Outputs (tables, figures) are transient and should be easily regenerated from data + scripts.

Look familar?

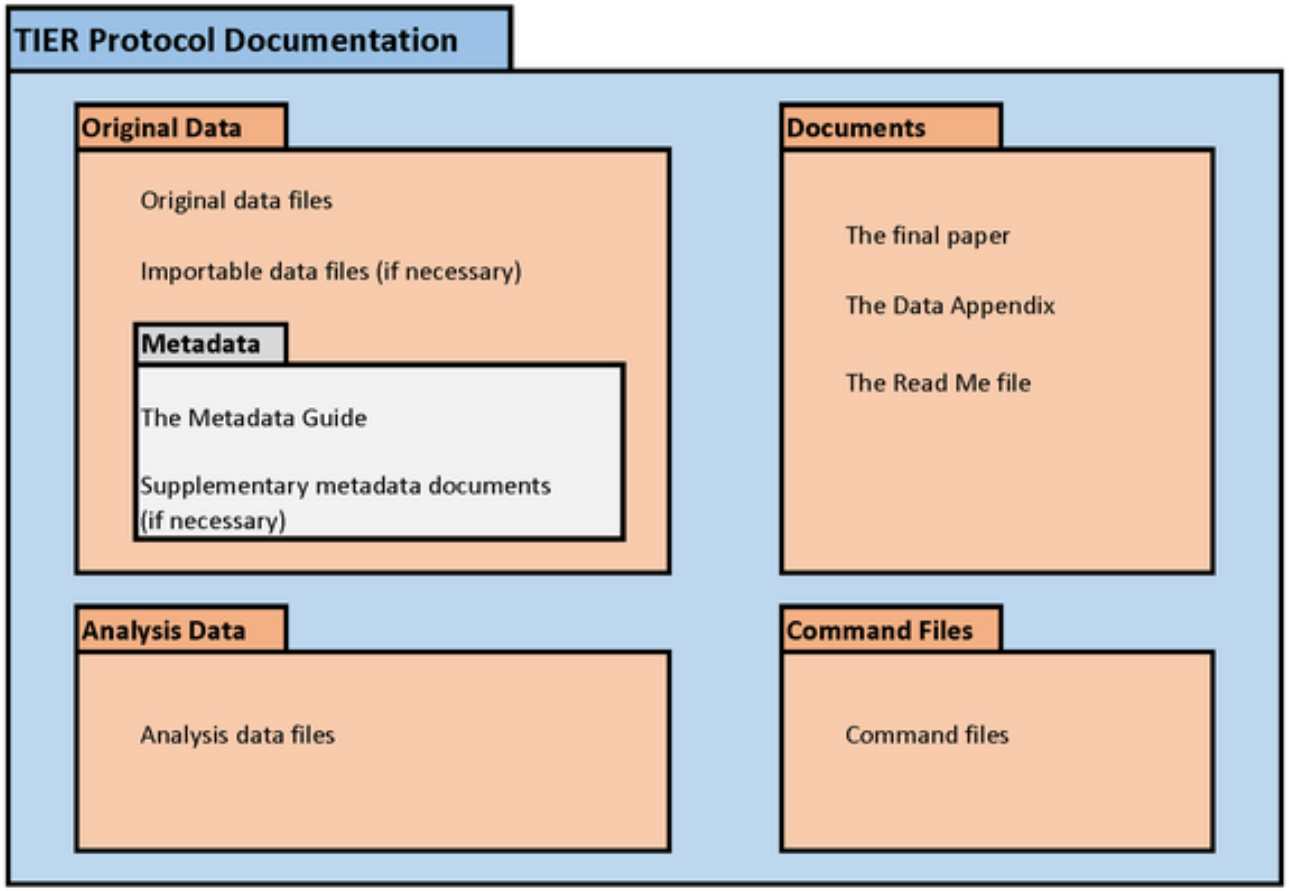
That's more like it.



Many variations (Project TIER scheme)

Also consider including a folder or file for 'metadata'.

Helps computers find your research!



(https://projecttier.org)

Tip: Don't obsess about rigidly following someone else's workflow.

Important thing is to find a workflow that works for you.

Your turn (finally!): Create a project

- Use RStudio or Stata (or something else) to create a new project
- Call it 'myproject'

- Create folders for `data`, `code`, and `output`.

- Now close the program.

- Browse to where the program file exists and double-click.
- Where is the current directory located? (Hint: type `getwd()` into the R console or `pwd` in the Stata command line).

- Are you in the right location? ✅ if yes, 🛑 if no

04:00

Can also be embedded in the context of a much larger project structure.

Note that "\Source" and "\Derived" data are never in the same folder.

Scott Long - Workflow Chapter 2 - designing a directory structure				
2008-04-17				
Project Directory	**Level 1**	**Level 2**	**Level 3**	**Purpose**
\AgeDisc				Project directory.
	\- To file			Files to examine and move to appropriate location.
	\Administration			Administration.
		\Budget		Budget sheets.
		\Correspondence		Letters and e-mails.
		\Proposal		Grant proposal and related materials.
	\Documentation			Documentation for project.
		\Codebooks		Codebooks for source and constructed variables.
	\Hold then delete			Delete when project is complete.
		\2007-06-13 submited		Do, data and text when paper was submitted.
		\2008-04-17 revised		Do, data and text when revisions are sumbitted.
		\2008-01-02 accepted		Do, data and text when paper is accepted.
	\Posted			Completed files that cannot be changed.
		\- Datasets		Datasets.
			\Derived	Dataset constructed from original data files.
			\Source	Original data without modifications.
		\- Text		Completed drafts of paper.
		\DataClean		Data cleaning and variable construction.
		\DescStats		Descriptive statistics and sample selection.
		\Figures		Graphs of data.
		\PanelModels		Panel models for discrimination.
	\Readings			Articles related to project; bibliography.
	\Work			Work directory.
		\- To do		Work that hasn't been started.
		\Text		Active drafts of paper.

Example from Long (2009)

3. Analytic Solutions

3.1 Workflow Management

3.2 Documentation

3.3 Literate Programming

3.4 Version Control

3.4 Dynamic Documents

"It is always faster to document it today than tomorrow."

-J Scott Long

What should be documented?

- Data sources
- Data decisions
- Statistical analysis
- Software
- Storage

Levels of documentation

- Research log
- Codebooks
- Dataset documentation

Research Log

The research log should help with:

- Keeping the work on track. Ideas, decision about analysis, papers you read that are relevant...document your thinking!
- Dealing with interruptions.
- Facilitating replication by others (especially by you 6 months later).

Ideally this should document what you did, why you did it, and how you did it.

Alexander Graham Bell did it, so can you!

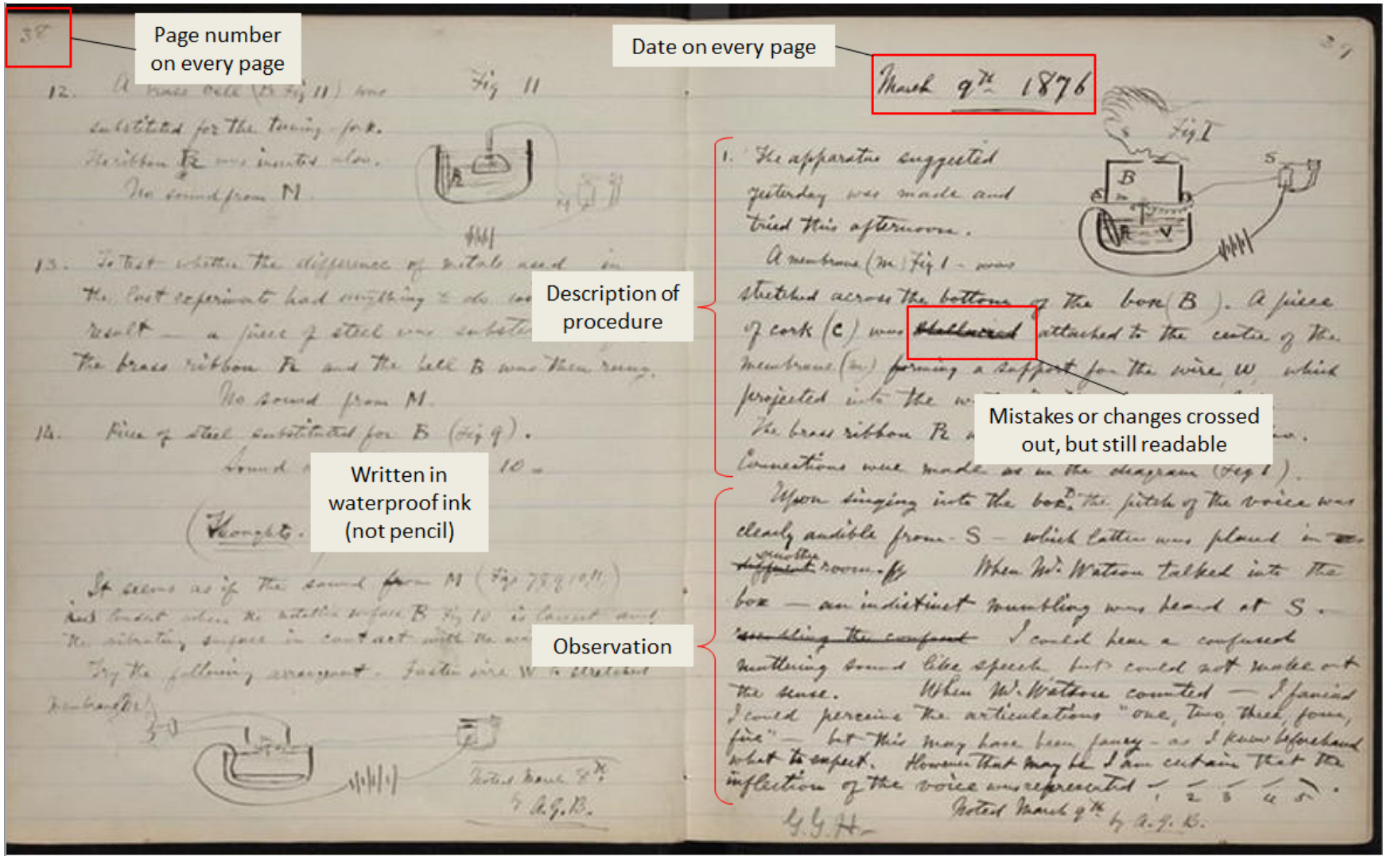


Image credit

Example from a (poor) research log

Any format is fine (text, Word, markdown). Requires discipline, but pays dividends later.

```
### 13 May 2018

* Set up as r-project. Collapsed all FARS data for each year 1975-2016 to zipped file, to be extracted during Stata run to
create analytic dataset, then deleted.
* Need to integrate data management with Stata but cannot figure out how to run a more complex .do file, so must resort to
running this bit in Stata and doing the rest of the analysis in R
* However, the package *RStata* does seem useful for running Stata from *within* R, which may prove very useful. A simple
example below:
```` {r RStata example}
library(RStata)
options("RStata.StataVersion" = 14)
options("RStata.StataPath"= '/Applications/Stata/StataMP.app/Contents/MacOS/stata-mp')
x <- data.frame(case = c(1356,2444), exp = c(1,0), pt = c(1,2))
stata("ir case exp pt", data.in = x)
````


14 May 2018

* Still not quite sure what to do about "special days" that are celebrated on different dates each year (Memorial Day, Labor
Day, Thanksgiving). Could potentially recode these and include as separate coefficients, similar to what can be seen in the
code [here](http://research.cs.aalto.fi/pml/software/gpstuff/demo_births.shtml) by collaborators of Gelman and the "birthday
problem" modeling births across days of the year:
```

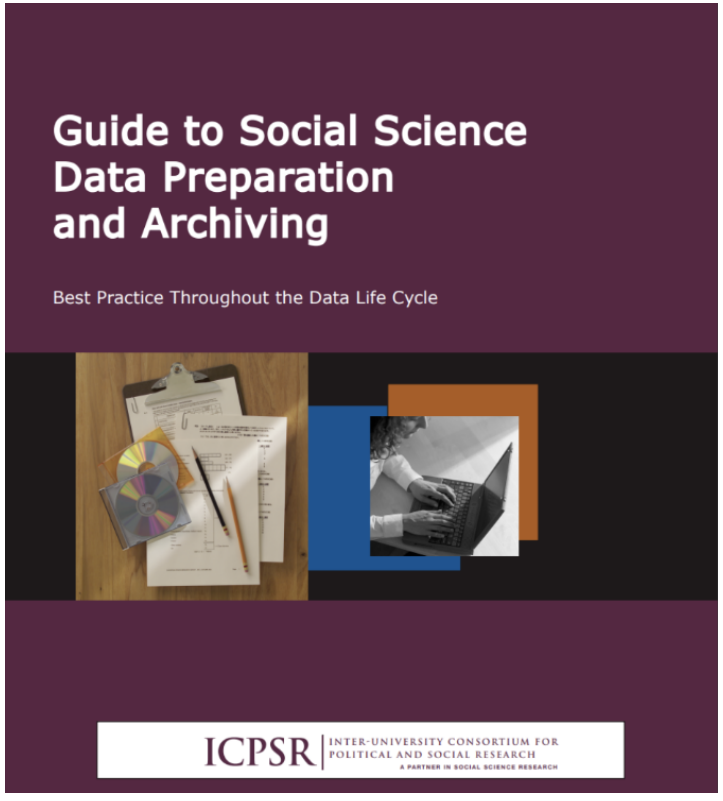
Codebooks

Codebook is crucial, especially if you are collecting your own data.

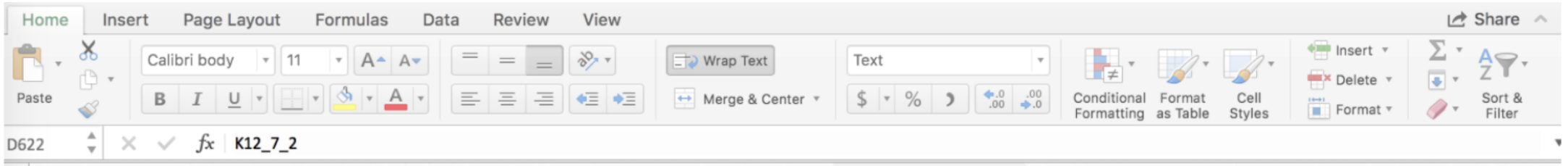
Should include:

- Variable name and question number.
- Survey questions.
- Skip patterns.
- Descriptive statistics.
- Missing values and reasons.
- Data imputation.

Good advice 👇



Codebook example (Excel)



	A	B	C	D	E	F	
1	**WAVE**	**QUESTION**	**QUESTION_TEXT**	**VARIABLE NAME**	**SECTION**	**POSSIBLE RESPONSES**	**NOTES**
2	*Wave number*	*Question number from survey*	*Exact text from the survey*	*Variable name from dataset*	*Section in the survey*	*List of response options*	*Please here, a (YYYY*
3	1	n/a	n/a	Interviewer_Name		n/a	RR 13 some a
4	1	A.1	Name of respondent	A1_1	Household characteristics	N/A	
5	1	A.2	Respondent ID number	id	Household characteristics	Integer, assigned to each respondent	AN 201 respon is actua Since A each ch househ only sa
6	1	A.3	Birthdate	A3_1_1 A3_2_1 A3_3_1	Household characteristics	Continuous date, month, and year	RR 201 variable reporte
7	1	A.4	How old are you?	A4_1	Household characteristics	Continuous number of years	RR 201
8	1	A.5	Have you ever attended school?	A5_1	Household characteristics	(1) Yes (2) No	

Codebooks can also be automated (using R 📦 `codebook`)

Import dataset: `codebook_data <- rio::import("fakedata.dta")`
Generate codebook: `codebook(codebook_data)` 👇

Codebook table

Copy CSV Excel PDF Print

Search:

name	label	type	type_options	data_type	value_labels	optional	scale_item_names	item_order	n_missing	complete_rate
session				character					0	1
created	user first opened survey			POSIXct					0	1
modified	user last edited survey			POSIXct					0	1
ended	user finished survey			POSIXct					0	1
expired				logical					28	0

Screenshot

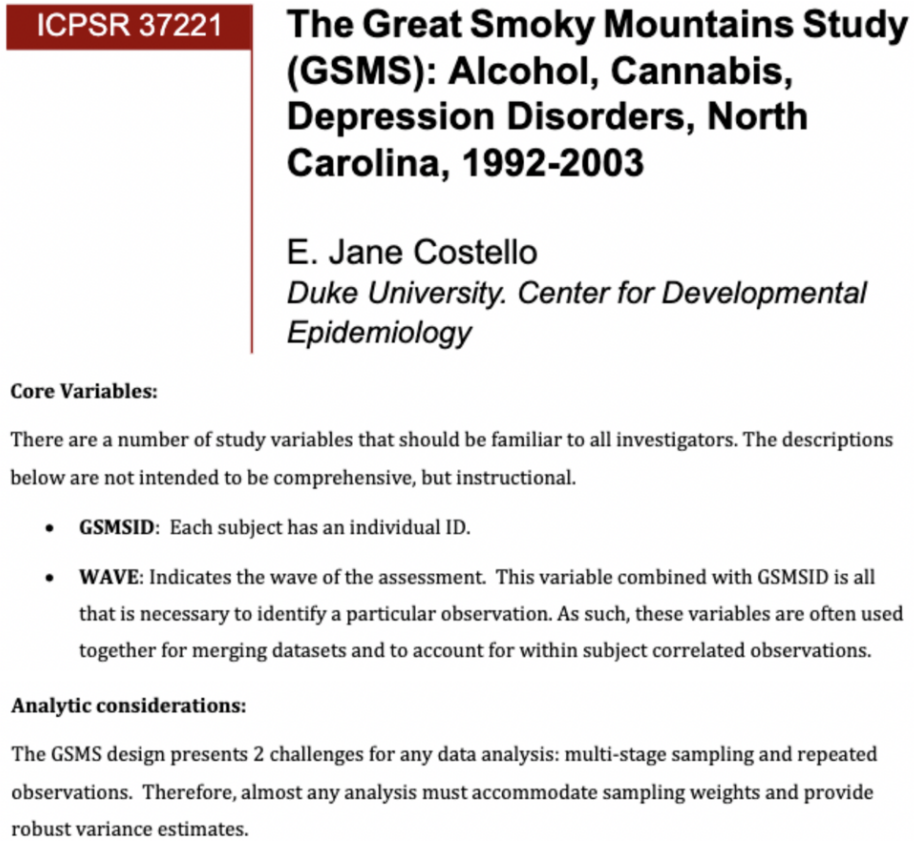
Dataset documentation

Why?

Provides key details about who, what, where, and which processes led to the created datasets.

How?

Ideally, using a reproducible format that can be updated regularly.



ICPSR 37221

The Great Smoky Mountains Study (GSMS): Alcohol, Cannabis, Depression Disorders, North Carolina, 1992-2003

E. Jane Costello
Duke University. Center for Developmental Epidemiology

Core Variables:

There are a number of study variables that should be familiar to all investigators. The descriptions below are not intended to be comprehensive, but instructional.

- **GSMSID**: Each subject has an individual ID.
- **WAVE**: Indicates the wave of the assessment. This variable combined with GSMSID is all that is necessary to identify a particular observation. As such, these variables are often used together for merging datasets and to account for within subject correlated observations.

Analytic considerations:

The GSMS design presents 2 challenges for any data analysis: multi-stage sampling and repeated observations. Therefore, almost any analysis must accommodate sampling weights and provide robust variance estimates.

One simple way: use a registry

Create a directory containing meta-data on each dataset created for the project.

Links directly to time-stamped source files that create data.

Dataset registry for `myproject`
Created by: NAME1
Last modified by: NAME2

Dataset	Filename	Date	Source	Comments
1	wave1.dta	2020-09-01	01-build-w1.do	Fix missing
2	wave2.dta	2020-10-05	02-build-w2.do	Ready to merge
...	...	...	...	...

3. Analytic Solutions

3.1 Workflow Management

3.2 Documentation

3.3 Literate Programming

3.4 Version Control

3.4 Dynamic Documents

"Literate" programming is:

Code that humans understand.

Computers will get it.

Core ideas for coding practices

Don't modify data by hand

This means no Excel for analysis, and no copy/paste.

Don't point-and-click

Typing commands or graphical interfaces are not reproducible.

Write code scripts

Coding creates reproducible workflow.

More general coding advice

Comment extensively!

Short lines with no wrapping

Give functions, objects, and variables intuitive names:

- `edu_percent` 👍
- `v76` 👎

Label variables and values (remember codebook advice).

Avoid abbreviations for commands (Stata).

Consider breaking long files into several shorter ones.

Essential parts of a syntax file

This example is for Stata, but the same principles apply to any software.

```
capture log close
log using _name_, replace text

//  program:    _name_.do
//  task:
//  project:
//  author:     _who_ \ _date_

//  #0
//  program setup

version 14
clear all
set linesize 80
macro drop _all

//  #1
//  describe task 1

//  #2
//  describe task 2

log close
exit
```

Generates a record of the session

Purpose, project, who made this file

Note version, since algorithms change

Better looking log

What is this bit of code doing?

And this one?

Same idea for SAS

01-processing.sas

CODE LOG RESULTS

```
/*
THIS DO-FILE OPENS THE DATA FROM THE
HARVARD SCHOOL OF PUBLIC HEALTH COLLEGE
ALCOHOL STUDY (ICPSR 4291),

THEN PROCESSES THE DATA TO PREPARE THEM
FOR ANALYSIS,

THEN SAVES THE PROCESSED DATA IN A FILE
CALLED analysis.dta

*WHEN YOU RUN THIS DO-FILE, MAKE SURE THAT
**********1) SAS's WORKING DIRECTORY IS SET
**********TO THE "Command-Files" FOLDER
**********2) THE HARVARD ALCOHOL STUDY DATA FILE
**********04291-0001-Data.dta IS SAVED IN THE
**********"Original-Data" FOLDER
*/

libname original  "/home/sam.harper/Original-Data";
libname analysis "/home/sam.harper/Analysis-Data";
filename ogdata "/home/sam.harper/Original-Data/04291-0001-Data.stc";

/* #1
  Read in the original dataset */

*import the original data;
proc cimport
  infile = ogdata
  library =  work
  ISFILEUTF8 = TRUE;
run;
```

Similar for R, all are adaptable

When commenting *your* code, consider how you feel when you read someone *else's* code!

```
# PROJECT
  # Paper:
  # Authors:

# R Script
  # Purpose:
  # Created: <date> by <author>
  # Updated: <date>
  # Inputs: <files required>
  # Outputs: <tables and figures>

######## PACKAGES ########
# Check system for packages
  need <- c("dplyr", "foreign") # list packages
```

```
######## ANALYSIS ########
# First fit crude model
fit1 <- lm(...)

# Now add sets of confounders
fit1 <- lm(...)

# Non-linear model for binary outcome
fit3 <- glm(...)

######## ROBUSTNESS ########
# Exposure measured with error (95% Se)
# Quantitative bias analysis using
# parameters from Smith et al. (2014)
fit_corrected <- lm(...)
```

Alignment, indentation, automation are your friends...

Harder to read

```
keep sid class_id teacher grade1 ///
grade2 grade3 pass
```

Easier to read

```
keep   sid class_id teacher grade1 ///
       grade2 grade3 pass
```

Without automation

```
//  create binary variables without loop
generate y_lt2 = y<2 if !missing(y)
generate y_lt3 = y<3 if !missing(y)
generate y_lt4 = y<4 if !missing(y)
```

With automation

```
//  create binary variables using a loop
foreach cutpt in 2 3 4 {
    gen y_lt`cutpt' = y<`cutpt' if !missing(y)
  }
```

Save outputs as objects to insert into papers

R:

```r
# a boring regression
lm(dist ~ 1 + speed, data = cars)

# save regression results for later
fit <- lm(dist ~ 1 + speed, data = cars)
```

Note that `fit` contains our coefficients and SEs. Now write to a table:

```r
library(broom) # need broom pkg
write.csv( tidy( fit ), "t1.csv")
```

Now `t1.csv` has estimates.

Stata:

```
webuse auto, clear

// another boring regression
regress price mpg
estimates store m1
```

Here `m1` contains our model results. Write to a table:

```
// need eststo package
esttab m1 using t1.csv
```

Now `t1.csv` has estimates.

Can be adapted to other software

Key idea:

- Results are *static* tables.
- Inserted into compiled manuscript.
- No copy/paste.

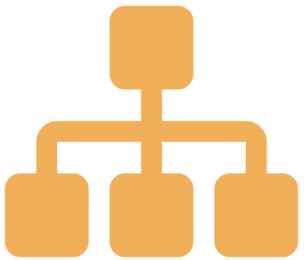
Example using SAS's **O**utput **D**elivery **S**ystem (ODS):

```
data etoh;
set analysis.analysis;
run;

ods select ParameterEstimates;
ods csv file = "/home/sam.harper/Output/model.csv";
proc logistic data=etoh;
model drunk = free;
run;
ods csv close;
```

Results written to "model.csv" file

Structure and name files intuitively.



Give code, data files, and output logical names where possible.

Number scripts sequentially in the order they should be run, e.g.,:

- 01_clean_data.R (or .do, .sas, etc.)
- 02_main_analysis.R
- 03_robust_checks.R

Label output figures with descriptive names, but ones that aren't likely to change (e.g., figure_hte.png is better than figure_1.png)

See Jenny Bryan's talk on naming things.

"One script to rule them all"

Many papers require multiple scripts.

Dependencies create problems.

Create a master-script to run everything.

Facilitates '1-click' reproduction.

Example of a "master" script (Stata)

This is the content of `code/mvc-laws-master.do`:

```
capture log close master
log using "code/mvc-laws-master", name(master) replace text

//  program:  mvc-laws-master.do
//  task:     run all analyses
//  project:  mandatory seat belt laws and traffic deaths
//  author:   sam harper \ 23feb2017

// Required user-written programs: -tabout, -metan, -estout, -egenmore
// Can be downloaded by typing "ssc install tabout, replace"  and
// "ssc install metan, replace", etc. into the Stata command line


do "code/mvc-laws-data.do"                    // create analytic dataset
do "code/mvc-laws-analysis-descriptives.do"   // descriptive table
do "code/mvc-laws-analysis-models.do"         // model results
do "code/mvc-laws-analysis-appendix.do"       // appendix tables/figures

log close master
exit
```

Your turn #2: Let's create a master script.

1) Open the `myproject` you created in R or Stata and create a new script (File>New File> R Script or File> New > Do-file. Type the following code:

- R: `print("Hello world!")`
- Stata: `disp "Hello world!"`

2) Save it as `1-hello.R` or `1-hello.do` in the "code" folder.

3) Repeat, creating another new script for `2-goodbye.R`, but type `"Goodbye world!"` instead of `"Hello world!"`

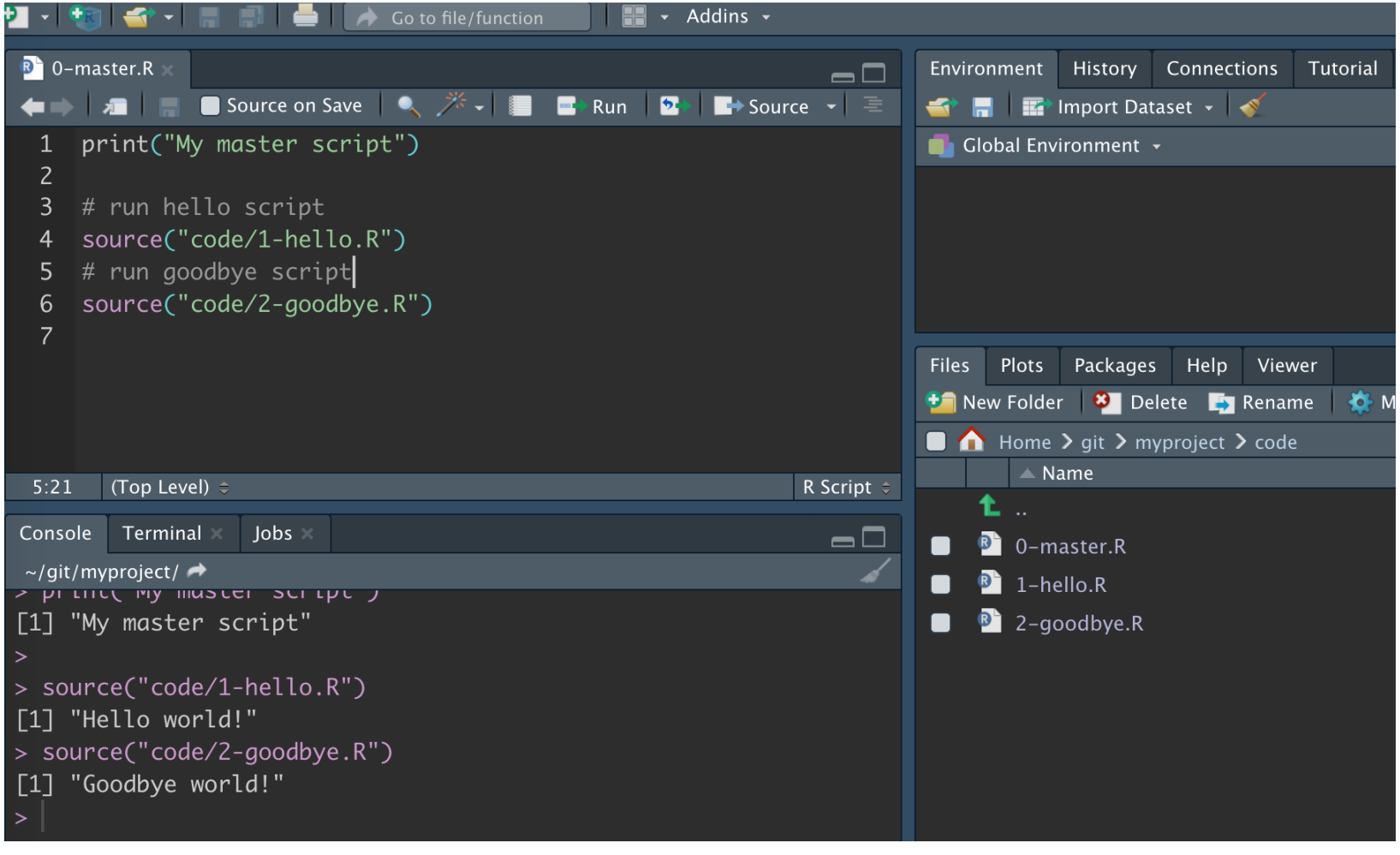
Now create a master script called `0-master.R` or `0-master.do` that will run the two other scripts you created. The command to execute another file is `source` in R and is `do` in Stata.

4) Run the master script.

- Success? ✅ if yes, 🛑 if no

05:00

Something like?



Break! ☕

10:00

3. Analytic Solutions

3.1 Workflow Management

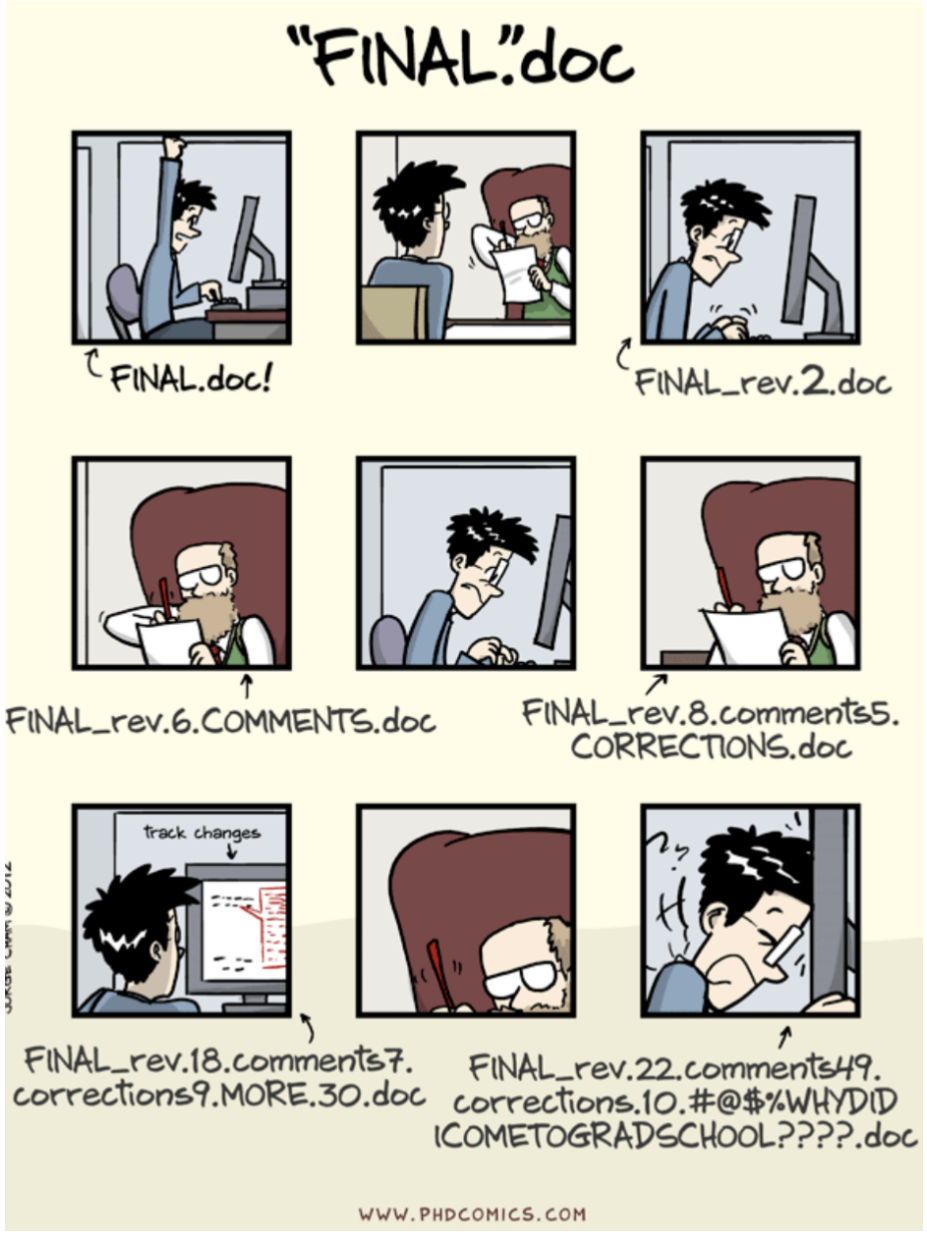
3.2 Documentation

3.3 Literate Programming

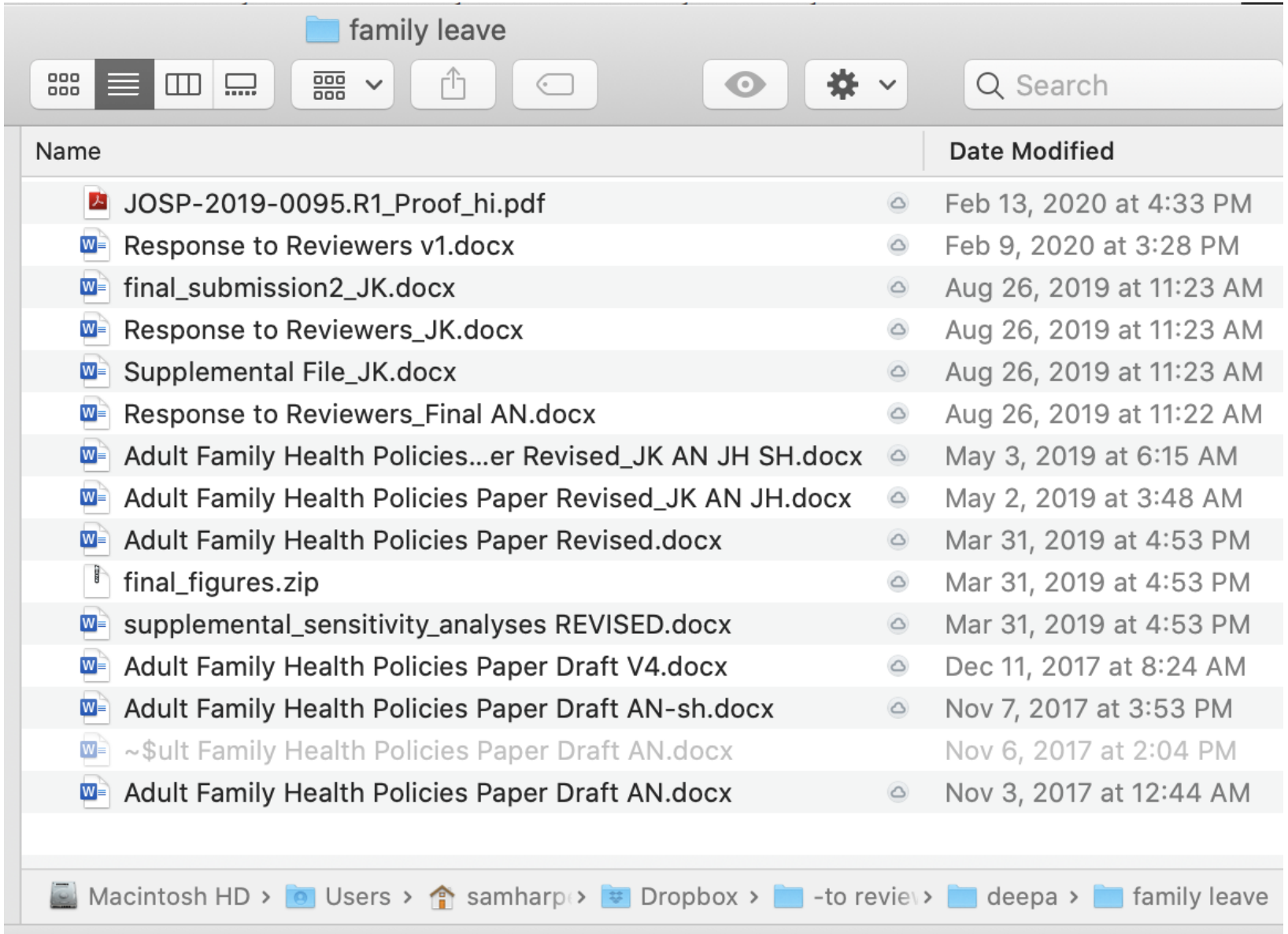
3.4 Version Control

3.4 Dynamic Documents

Obligatory picture for every discussion of version control.



But not a joke 😞



One kind of collaborative workflow

Collaborator 1:

- cleans data → `data.do`
- analysis → `analysis.do`
- generates Excel tables → `results.xls`
- draft manuscript in Word → `manuscript.docx`

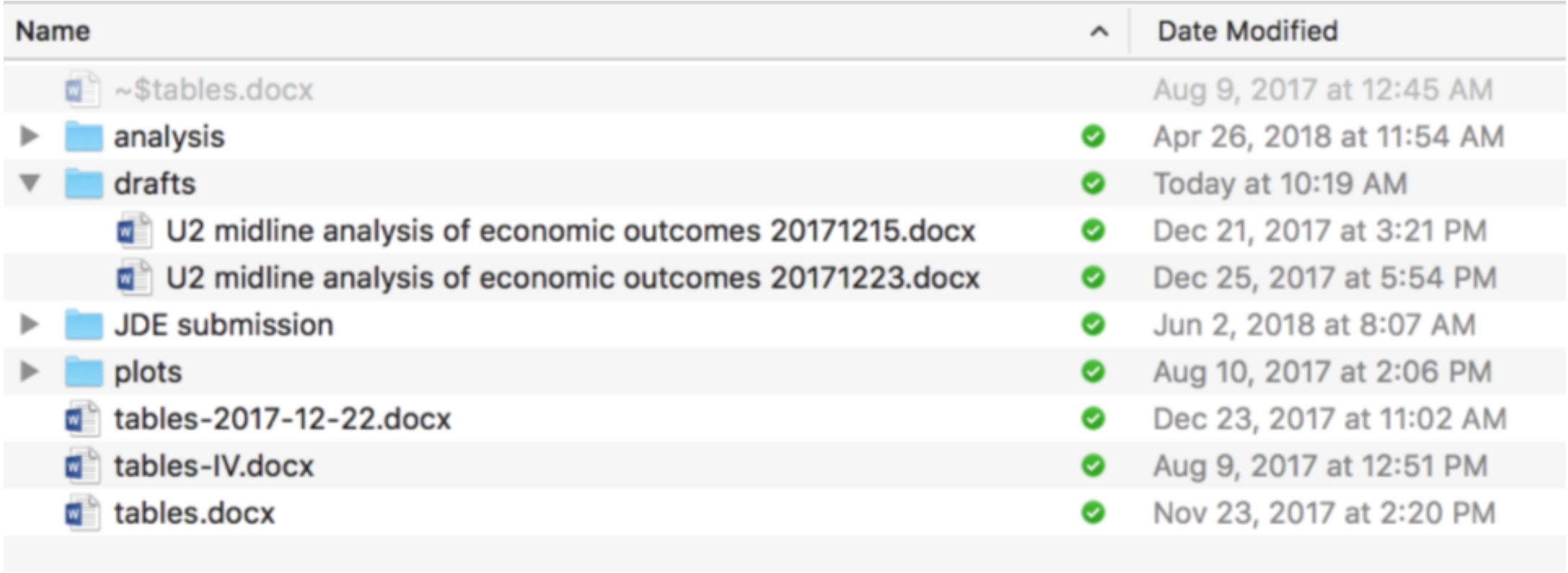
Collaborator 2:

- adds new variables → `data_new.do`
- new analysis → `analysis2.do`
- new tables → `results_final.xls`
- revises draft with new results → `manuscript-v2.docx`

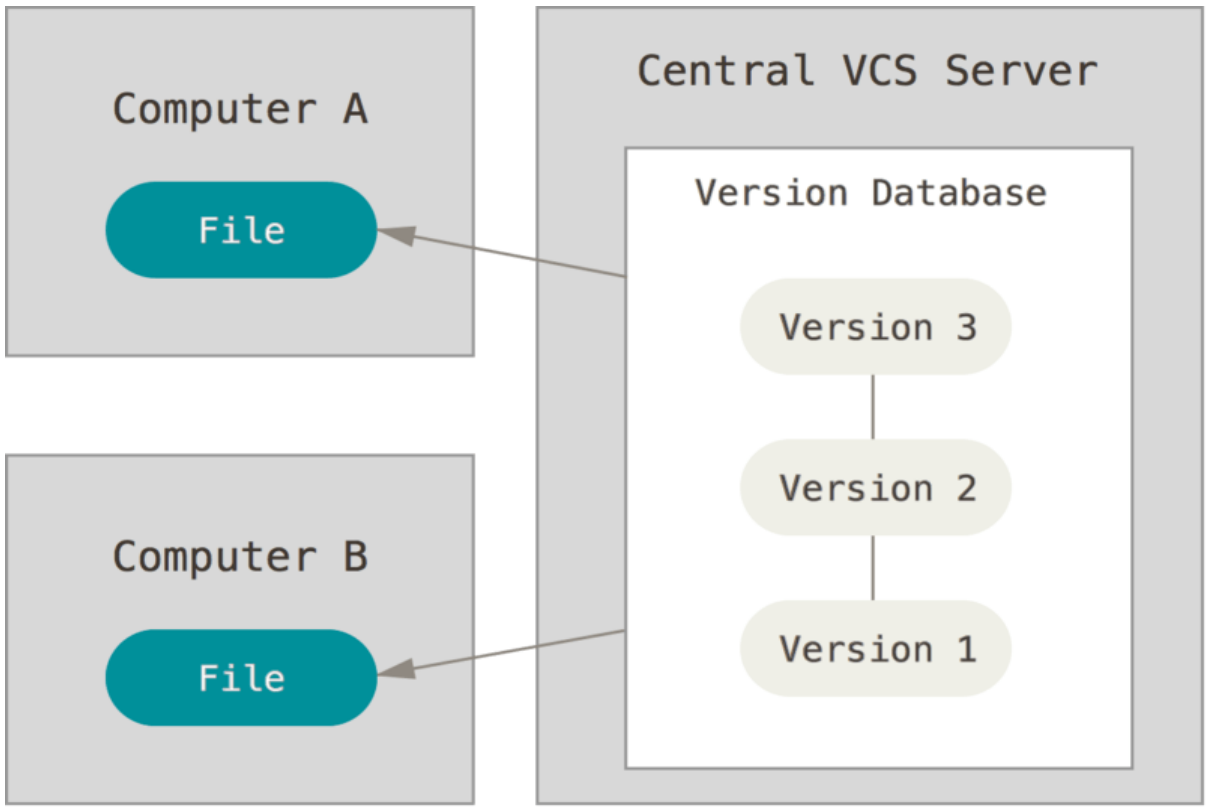
This can work! But it can also create problems.

What is the problem?

Which draft is the current one? Which set of tables contain the "right" results?



A version control system (VCS) automatically keeps track of changes to files and code.



How can version control help?

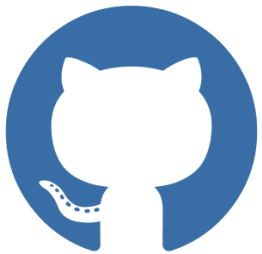
Under VCS the prior collaboration would only have 4 files:

- cleans data → data
- analysis → analysis
- generates results → results
- generates manuscript → manuscript

A VCS keeps track of changes to each file.

We just don't usually see it.

How can version control help?



With version control you can:

- Collaborate
- Track each change and **who** made it
- Easily switch between versions of files (i.e. go backward in time)
- Compare versions of files
- Backup
- Work with the same files on different machines
- Collaborators can work simultaneously on the same files on different machines
- Experiment with a new version of code without permanently ruining anything

Manual solutions (works, but not ideal)

Create dated versions of files (save-as) for each substantive change

With each modification, re-run ALL code to make sure nothing is broken—helps if you have a master file to run all scripts!

Check-in with coauthors to ensure multiple people aren't working on the same files at the same time.

Keep a simple log to remind yourself of the location/content of major changes.

Software solutions

Version control software: helps manage versions and edits to files.

Git: Open-source, “distributed model” of version control developed by creator of Linux.

GitHub / GitLab: Free, web-based service that hosts Git “repositories” and offers a variety of features for collaboration:

- Online
- Desktop apps
- Command line (more technical)

How does Git work?

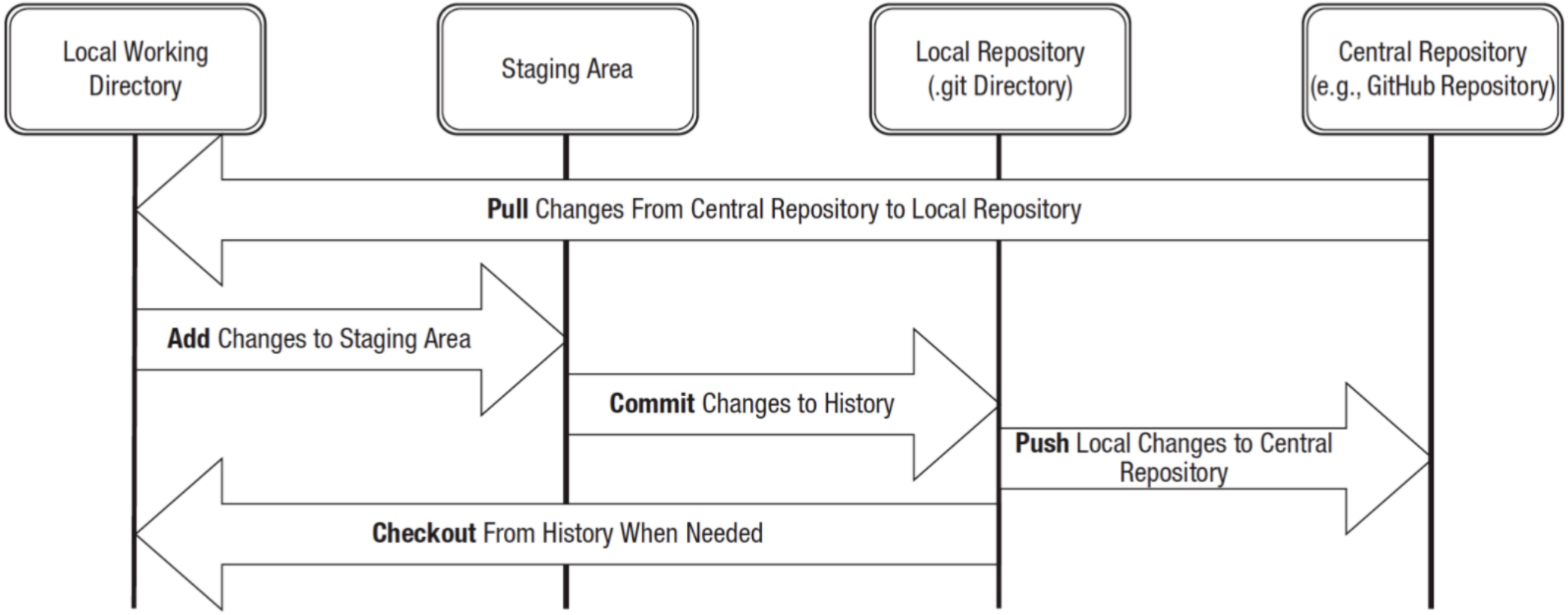


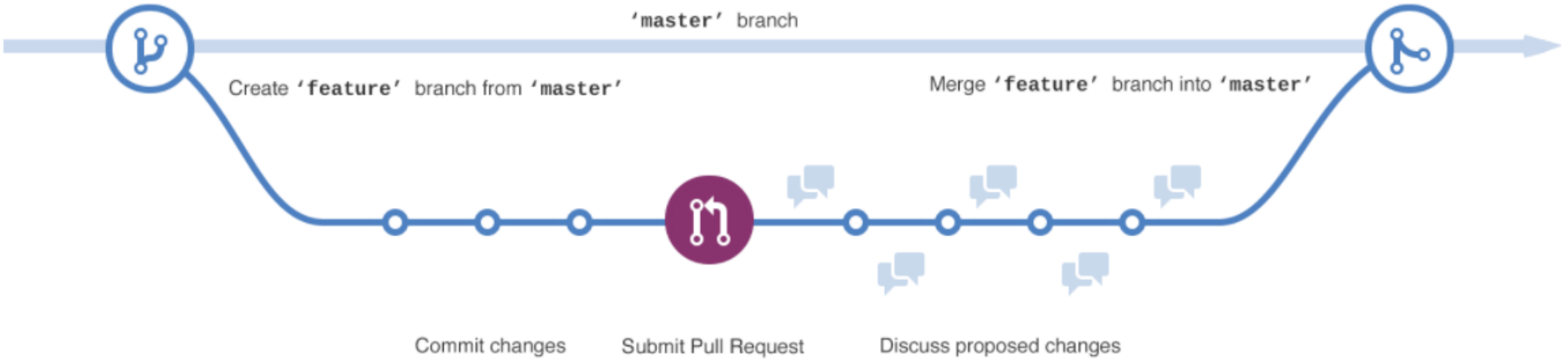
Fig. 9. A diagram illustrating the typical collaborative Git workflow with a central repository (e.g., GitHub). As when a central repository is not involved, users work in their local repositories, making changes to files and viewing prior versions as needed. In addition, they push changes from their local repositories to the central repository, so that collaborators have access to them, and pull changes from the central repository to their local repositories, to see their collaborators' work. Verbs in boldface are Git operations and explained in the main text.

Advantages of Git

Provides the entire narrative of your project.

Allows you to go back in time.

Experiment without breaking things.

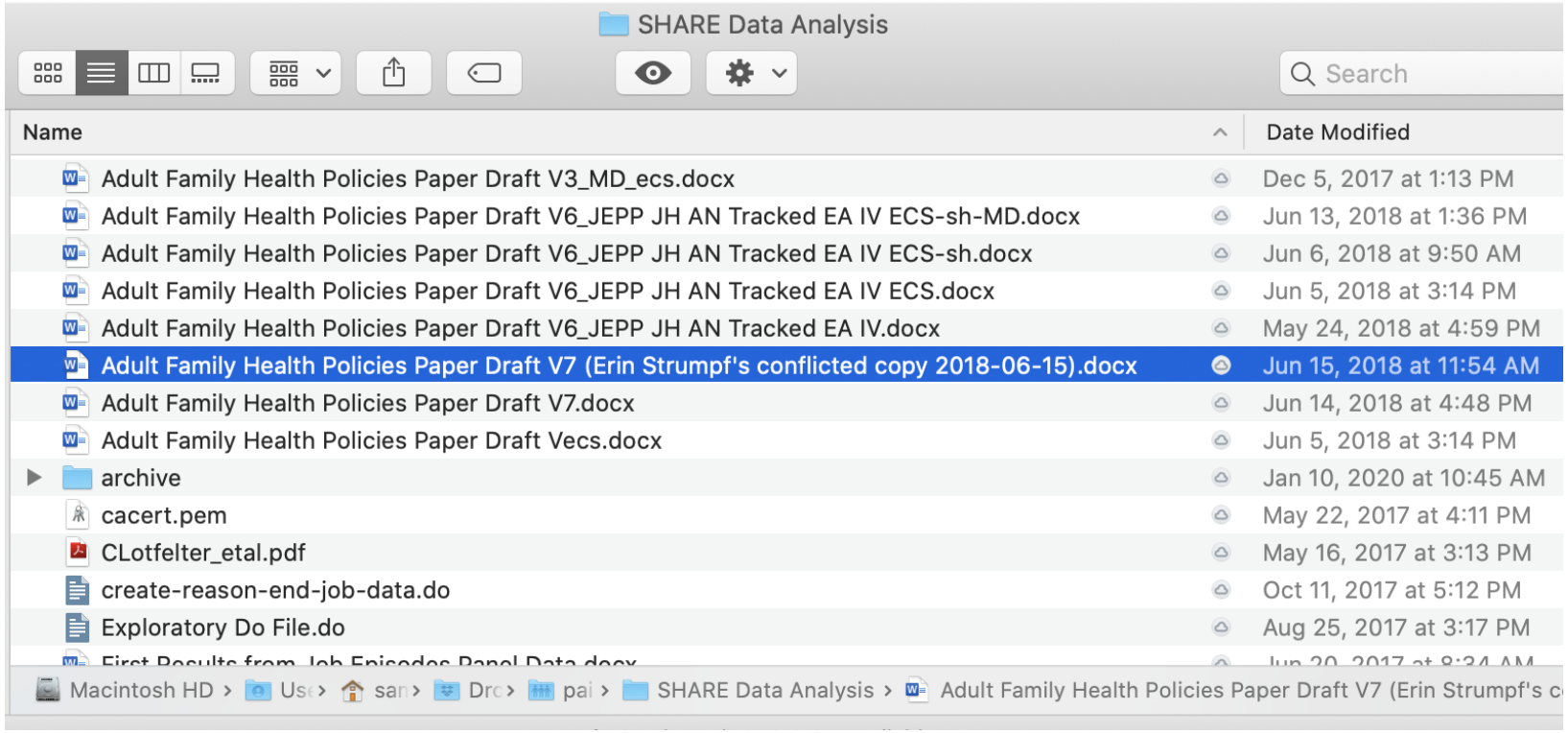


How is that better than Dropbox?

Simultaneous editing in Dropbox leads to "conflicted" copies of files.

Only one person can edit the "live" version of the file.

Can "rollback" to earlier version, but that affects everyone.



Additional resources for Git

- The Basic Workflow of Git (an infographic explaining how Git's version control system works): https://www.git-tower.com/blog/workflow-of-version-control
- Git + GitHub (information on using Git and GitHub in an R programming context): http://r-pkgs.had.co.nz/git.html
- GitHub's Git cheat sheets (reference sheets on the most commonly used Git commands)
- GitHub Glossary (a glossary of Git and GitHub terminology)
- Pro Git (Chacon & Straub, 2014; a complete manual of Git)
- tryGit (an interactive Web site for learning the basics of Git)
- Jenny Bryan's HappyGit

Vuorre & Curley (2018)

3. Analytic Solutions

3.1 Workflow Management

3.2 Documentation

3.3 Literate Programming

3.4 Version Control

3.4 Dynamic Documents

What are dynamic documents?

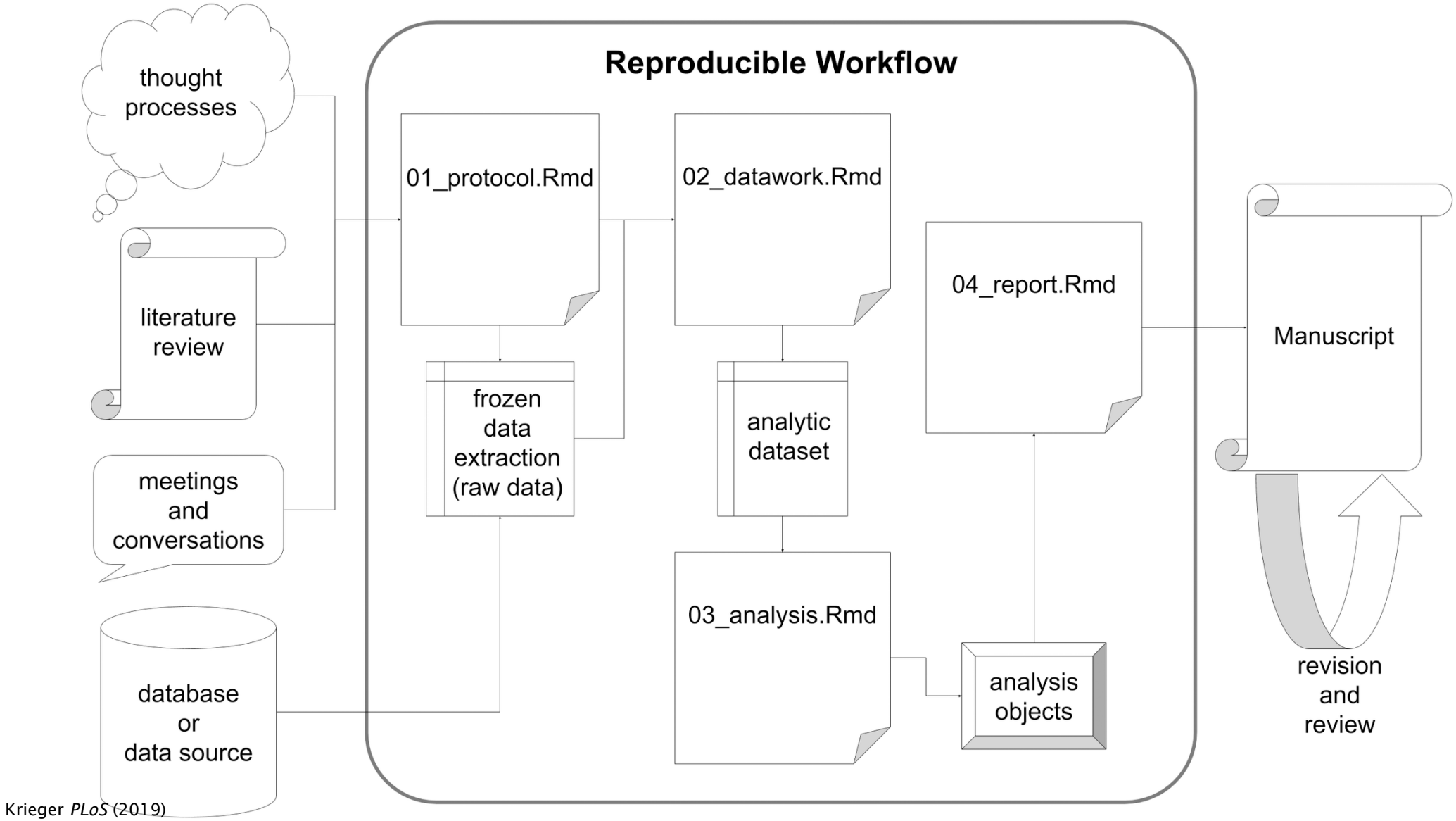
Even with perfect (version controlled) code, you can still run into problems going from your code to paper. (copy → paste problems)

This is where dynamic documents come in.

A dynamic document includes your data, code, analysis, and output all in one place.

Fully automated so no mistakes from copying and pasting.

Several formats, but most commonly using R Markdown in RStudio or Markdoc in Stata.



Krieger *PLoS* (2019)

How can dynamic documents help?

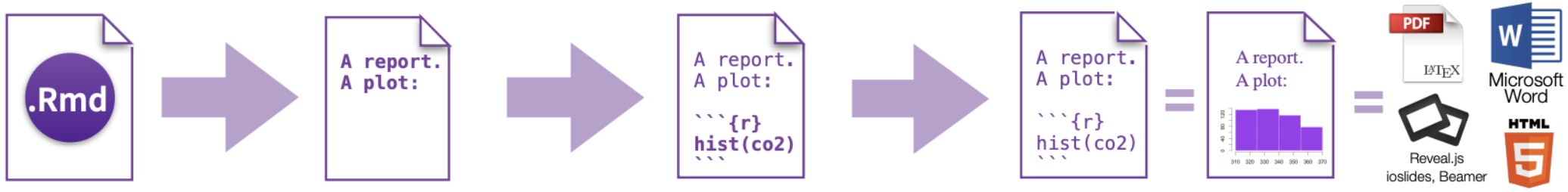
- Include tables by linking to a file, instead of a static image.
- Reference an estimate (e.g. risk ratio) by linking to a value calculated by an analysis file, instead of a static number typed manually.
- Automatically update tables and numbers if models and/ or data changes.
- Produce entire paper with one or two clicks.
- Most written using markdown language.

i. Open - Open a file that uses the .Rmd extension.

ii. Write - Write content with the easy to use R Markdown syntax

iii. Embed - Embed R code that creates output to include in the report

iv. Render - Replace R code with its output and transform the report into a slideshow, pdf, html or ms Word file.



	Markdown	Result

Headers	`# Header 1`
	`## Header 2`
	`### Header 3`
	`#### Header 4`
	`##### Header 5`
	`###### Header 6`
		Header 1
		Header 2
		Header 3
		Header 4
		Header 5
		Header 6

Text	`*italic* _italic`
	`**strong** __strong__`
	`***italic & strong*** ___i&s___`
	`~~strikethrough~~`
		italic
		strong
		italic and strong
		~~strikethrough~~

Links	`https://arminreiter.com`
	`[MyLink](https://arminreiter.com)`
	`[MyLink](https://arminreiter.com "Title")`
	`[MyLink][1]`
	`[MyLink][URL to AR]`
	`See my [link]`
	
	`[1]: https://arminreiter.com`
	`[URL to AR]: https://arminreiter.com`
	`[link]: https://arminreiter.com`
		https://arminreiter.com
		MyLink
		MyLink
		MyLink
		MyLink
		See my link

Images	`![Logo](/images/logo.png)`	

Example of a whole paper written and submitted with Rmarkdown

Stata catching up

Since Stata v15, now the `dyndoc` command can produce Word or HTML documents from Markdown.

More limited functionality than `markstat` but will likely improve.

Creating a **dyndoc** file is as easy as creating a do-file. Here's one that creates an HTML file with only Stata output:

```
──────────────────────────────────────────── example1.txt ────
~~~~
<<dd_do>>
webuse auto, clear
summarize price
<</dd_do>>
~~~~
──────────────────────────────────────────────────────────────
```

The four tildes in a row, **~~~~**, are Markdown syntax to start and end a code block.

Terms in **<<** … **>>** are called Stata dynamic-document tags. The code block is bounded by **<<dd_do >>** … **<</dd_do>>**. Stata code inside **<<dd_do>>** … **<</dd_do>>** is executed and its output substituted into the HTML document. We merely save the file above as **example1.txt**, type **dyndoc example1.txt** in Stata, and **example1.html** is created for us:

```
. webuse auto, clear
(1978 Automobile Data)

. summarize price

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
```

Main elements of a dynamic document

Header information

Document metadata (author, title, date, formatting options).

Document text

Write using markdown language (incl. equations, refs)

Code: both 'inline' and as 'chunks'

Performs data maniupation, analysis, tables, figures.

Header (also called YAML*)

Rmarkdown:

```
---
title: "My Reproducible Paper"
author: Sam Harper
date: 2020-10-30
{{output: pdf}}
bibliography: myrefs.bib
csl: vancouver.csl
---
```

Other output formats available (e.g., .html or .docx output), many other formatting options.

Same structure for Stata Markdown:

```
---
title: "My Reproducible Paper"
author: Sam Harper
date: 2020-10-30
abstract: |
  I give a brief summary here.
keywords: |
  markdown, reproducible research.
bibliography: myrefs.bib
---
```

Document format specified in another step.

*YAML stands for YAML Ain't Markup Language. Ha.

Document text

Rmarkdown:

```
---
header:
---
```

Text written using markdown language:

```
# Introduction
This is why this paper is necessary, and here
are the gaps it will fill.

# Methods
## Data
We used some cool data.
```

Same structure for Stata Markdown:

```
---
header:
---
```

Can add references or equations as well:

```
## Statistical analysis
We ran a linear regression [@Galton:1875]:

$$
y_{it}=\beta_{0} + \beta_{1}*X
$$

where $ \beta_{1} $ is what matters.
```

Adding inline or 'chunks' of code

Rmarkdown:

```
---
header:
---
```

Document text including code 'chunk', set apart by 3 backticks:

````
```r{eval=FALSE}
fit1 <- lm(y ~ x, data=d)
```
````

Inline code: single backtick `` `r 1+1` `` .

Structure for Stata Markdown:

```
---
header:
---
```

Same idea for code 'chunks':

````
```s
// Write stata code inside chunks
sum mpg
```
````

Inline: single backtick `` `s r(mean)` `` .

Allow text to update automatically when results change.

Rmarkdown:

Header + document text followed by code chunk:

````
```r{eval=FALSE}
fit1 <- lm(y ~ x, data=d)
est <- summary(fit1)$coefficients["x","Estimate
```
````

We find x increases y by `` `r est` `` .

Will update if data or estimates change.

Structure for Stata Markdown:

Header + text with Stata code chunks:

````
```s
reg price mpg
```
````

Mpg increases price by `` `s _b[mpg]` `` with a standard error of `` `s _se[mpg]` `` .

These are dynamic values.

Stata markdown example (start at project directory)

Stata do-file (`code/paper.do`) calls the markdown file using the `markstat` command:

```
// change to writing directory
cd writing


// generate the manuscript
markstat using "paper.stmd", pdf

// back to the main directory
cd ..
```

Markdown text (`writing/paper.stmd`)

````
---
title: My New Paper
author: Sam Harper
date: 2020-10-28
---

# Introduction

We did some stuff. And made a plot

```s/
    quietly webuse auto, clear
    quietly scatter weight length, legend(off)
    quietly graph export scatter.png, width(800) replace
    quietly sum price
    scalar mp = round(r(mean),1)
```

![Correlation between weight and length](scatter.png)

We found the mean price was `s mp`.
````

My New Paper

Sam Harper

2020-10-28

Introduction

We did some stuff. And made a plot

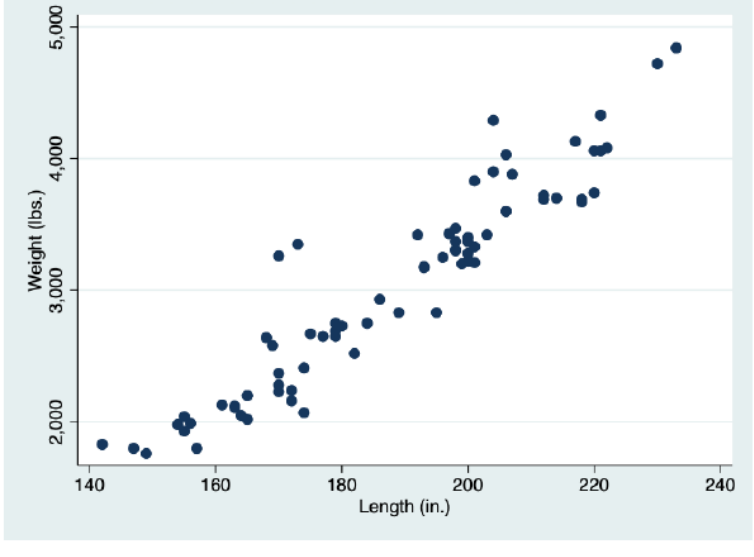


Figure 1: Correlation between weight and length

We found the mean price was 22.

Knit Insert Run

```
---
title: "My SAS Example"
author: "Sam Harper"
date: "26/10/2020"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(here)
library(kableExtra)
```

Results

We ran a logistic model:

```{r model, echo=F, message=FALSE}
t1 <- read_csv(here("slides/03-analytic", "model.csv"), skip=2,
  col_names=c("Parameter", "df", "Estimate", "SE", "X2", "pvalue"))
kable(t1, format = 'latex', digits=2,
      caption = "Logistic Estimates", booktabs = T) %>%
      kable_styling(latex_options = "hold_position")
```

Results look good!
```

Can be adapted to other software

Rmarkdown options, packages required

Here are your model results

Rmarkdown + results from SAS

1. Use SAS to fit models, generate results.
2. Export results to files.
3. 'Knit' together with text using Rmarkdown.

My SAS Example

Sam Harper

26/10/2020

Results

We ran a logistic model:

Table 1: Logistic Estimates

Parameter	df	Estimate	SE	X2	pvalue
Intercept	1	0.79	0.05	257.07	<.0001
free	1	0.00	0.09	0.00	0.9574

Results look good!

Break! ☕

10:00

Reproducible Research: Why and How

Part 4: Dissemination Solutions

Sam Harper

SER Pre-Conference Workshop
2020-10-30

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

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

Should:

- be complete but parsimonious. Don't over do it.
- run / reproduce results with minimal effort (1-click).
- be "literate" (human readable).
- protect confidential data.

There is no single, perfect way to organize or prepare files for replication.

Find a workflow that works well for you.

Clark (2017)

Step 1: Generate replication files

1. Create a new organized empty replication folder within your project directory (e.g., “replication_files/”)

2. Subfolders: Should be same as overall file structure:

    - code/ — scripts
    - data_clean/ — manipulated data
    - data_raw/ — original data
    - output/ — generated tables, graphs, etc.
    - extra/ — misc. extras (e.g., code book)

3. A "README" text/markdown file to document contents, sources, software/system versions, other info necessary for replication/comprehension.

Clark (2017)

Step 2: Replicate your own results

1. Copy data and code to your new replication directory.
2. After copying all of the relevant files, see if you can replicate the results in your paper.
3. May want to start with the "final" products (i.e., tables and figures from clean data), which should be "easiest" to replicate.
4. Check for errors and make sure all is well.
5. Now copy the original raw data and cleaning scripts and run the entire thing.
6. All good? If not, debug and try again.

Clark (2017)

Step 3: Final check

1. Shut down and restart software package.
2. Replicate again...all good?
3. Or have a friend / colleague try on another computer.
4. Fix any remaining bugs and try again.
5. Now ready to disseminate!

Clark (2017)

How replication can help

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.

Hatzenbuehler et al. *Soc Sci Med* 2014:

Reported 12 year decrease in life expectancy for sexual minorities living in more prejudiced communities.

Hatzenbuehler et al. *SSM* (2014). NY Times

Replication attempt using same public data "failed".

Re-analysis commissioned by original authors.

Coding errors discovered.

Study retracted.

> *This article has been retracted at the request of the authors and the Editors-in-Chief.*
>
> *The reason for the retraction is that the authors discovered an error in the study, which, once corrected, rendered the association between structural stigma and mortality risk no longer statistically significant in the sample of 914 sexual minorities. The authors published a Corrigendum (Corrigendum to "Structural stigma and all-cause mortality in sexual minority populations" [Soc. Sci. Med. 103 (2014) 33–41], Volume 200, March 2018, p 271), pending a re-analysis of the data. Re-analysis confirmed that the original finding was erroneous and the authors wish to fully retract their original study accordingly.*

Source: Retraction Watch

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

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.

What to share?

Pre publication

Preregistration/pre-analysis plan

Codebook / data documentation

Code to create / analyze data.

Replication files

Reports / preprints

Post publication

Peer-reviewed papers / postprints

Entire project?

What about pre-prints (or post-prints)?

Most publishers allow posting of a final "accepted" proof.

Consult the agreement you sign.

It is your work!

What rights do I retain as an Oxford Journal author?

- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

Standard Oxford University Press License (e.g., AJE, IJE)

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

Same code, different environment

Code may not run somewhere else. People are working on that:

Docker containers

packages up everything needed to run an application: code, runtime, system tools, system libraries and settings in a standalone, executable package.

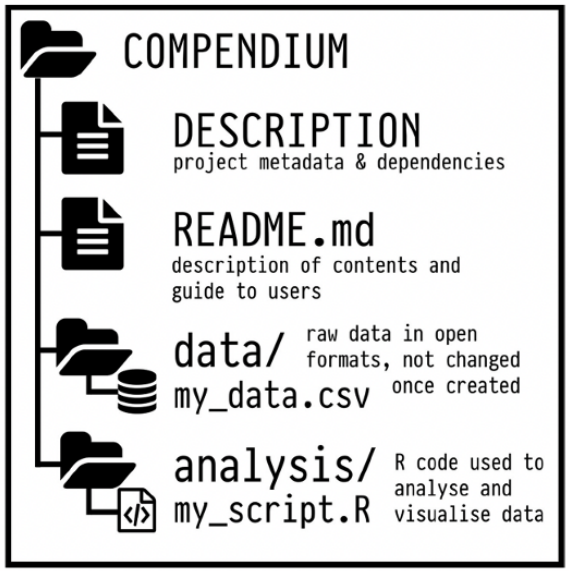
Binder

a service that provides your code and the hardware and software to execute it (mostly R, Python).

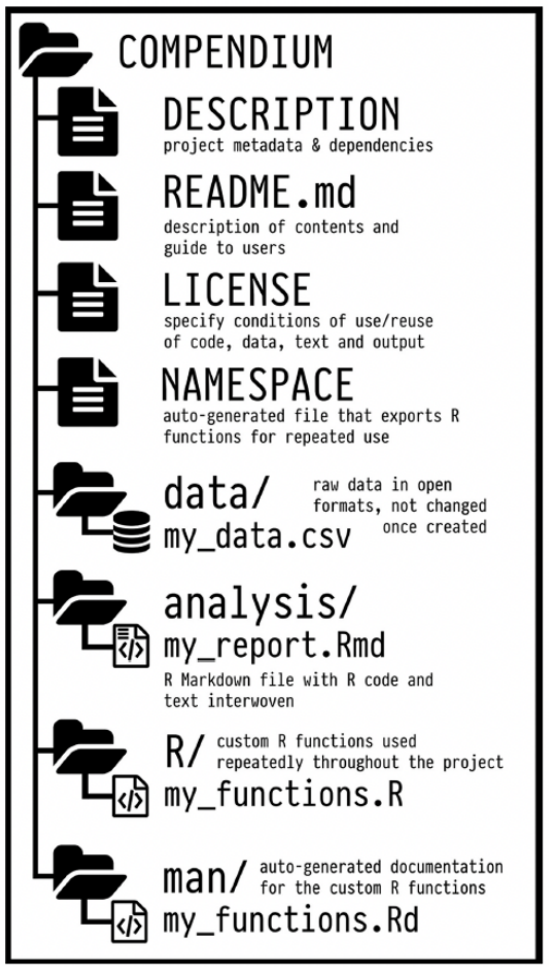
More advanced than we can cover today. See https://mybinder.org and https://hub.docker.com/ for more.

Can be done for any size project 👇

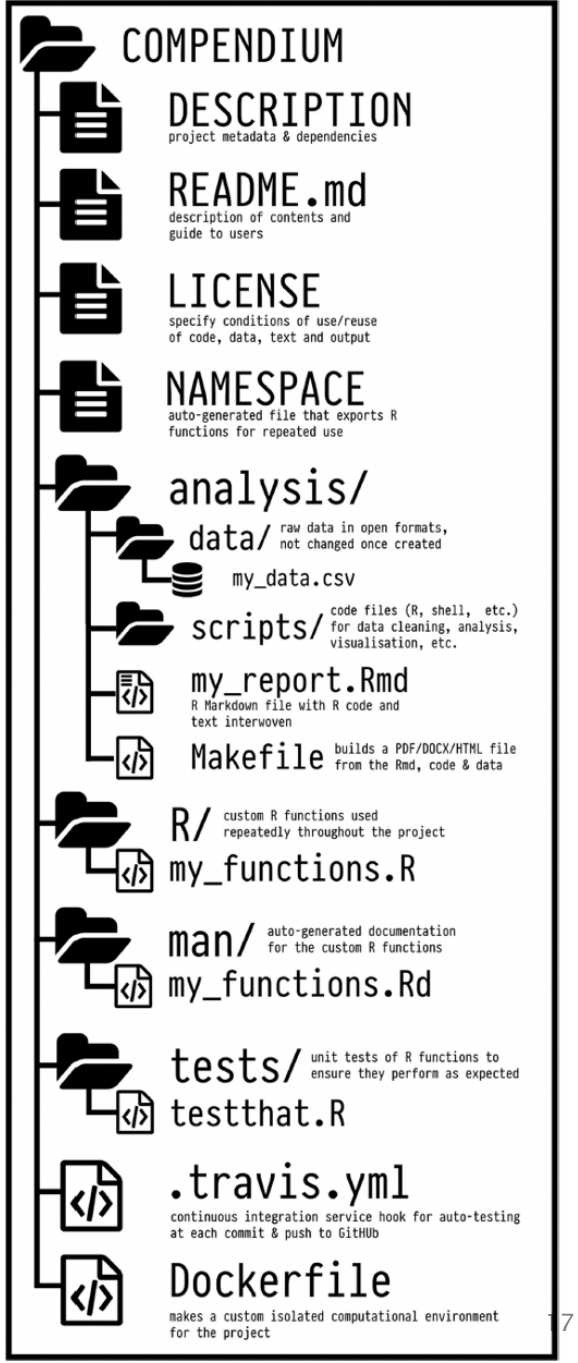
Small



Medium



Large



When to share data or code?

Many options

1. Include everything with submitted paper (public)

2. Include everything with submitted paper (for review)

3. Post-publication (recall TOP guidelines)

What if someone steals my idea?

Replication files may not matter!

Annals of Internal Medicine IDEAS AND OPINIONS

Dear Plagiarist: A Letter to a Peer Reviewer Who Stole and Published Our Manuscript as His Own

Michael Dansinger, MD

Dr. Doctor,

I am aware that you recently admitted to wrongly publishing, as your own, a scientific research paper that I had submitted to *Annals of Internal Medicine*. After serving as an external peer reviewer on our manuscript, you published that same manuscript in a different medical journal a few months later. You removed the names of the authors and the research site, replacing them with the names of your coauthors and your institution.

many research papers. It just doesn't make sense. Whether the pressure to publish is so intense, or whether the culture where you work is relatively permissive such that plagiarism is not taken as seriously, or whether getting caught seemed unlikely–it is hard to imagine why you would take this chance.

I hope you will not steal anyone else's research in the future. Instead, perhaps there is some way you can assist the scientific community's efforts to reverse the growing epidemic of plagiarism and scientific fraud.

Biased anecdote for benefits of sharing *with* paper submission

"Thanks for the opportunity to review this interesting paper. It is exciting to see the FARS data used in this way. **It was also exciting that you shared your code and this allowed me to review your work in a way I have not done before.**"

-Reviewer 1

"Overall verdict: This paper was both exciting and a pleasure to read. Clearly written, well argued, and with a highly commendable open science approach. As a referee, one frequently thinks 'why didn't they report the results of a model with...' - and in this case **I was able to download the data, assess their code scripts, and run my own specification to see how it changed things**."

-Reviewer 2

We probably got lucky, but the paper was accepted on the first submission.

Where to share?

Depends on discipline: find appropriate registry at http://www.re3data.org/, or check out ...

- Harvard's Dataverse
- Open Science Framework
- OpenICPSR
- figshare
- Data Dryad
- Many others...

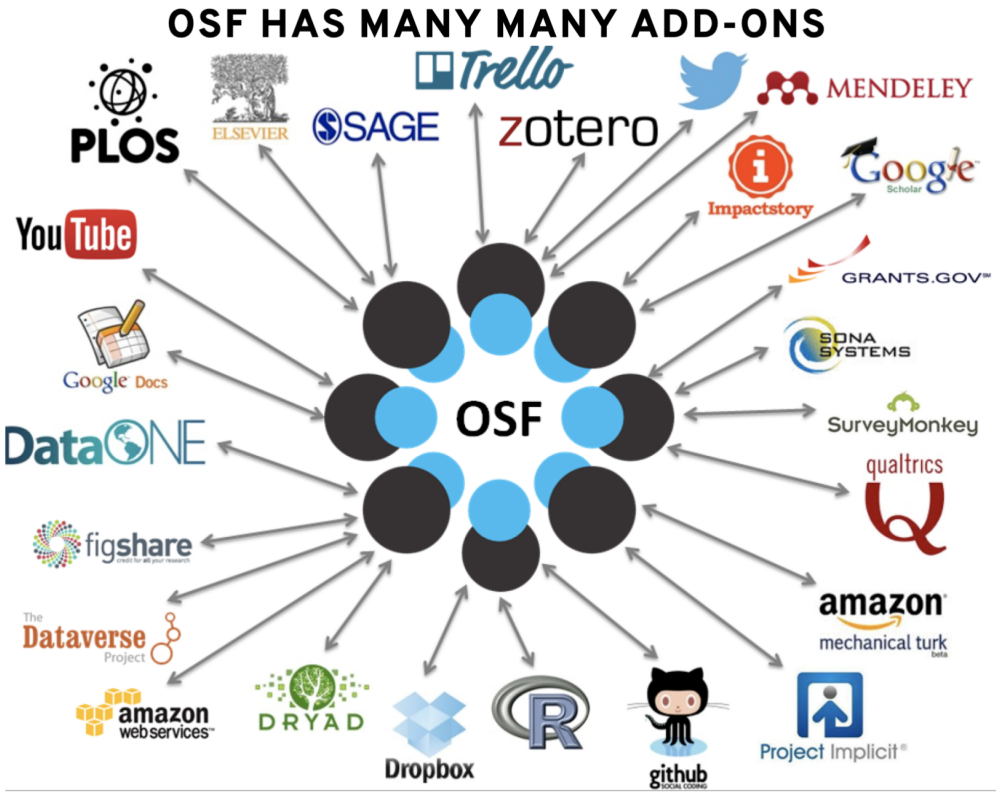
Open Science Foundation

OSF (http://osf.io) provides a central location to manage project files.

Any type of files can be uploaded (up to 5GB).

Most common file types will render to be viewed on OSF.

OSF also provides a more comprehensive system for planning, documenting, executing, and disseminating your research over the entire life cycle of a project--and beyond.



OSF workflow

1. Create a structured workspace.

2. (Possibly) pre-register study

3. Deposit / add study materials.

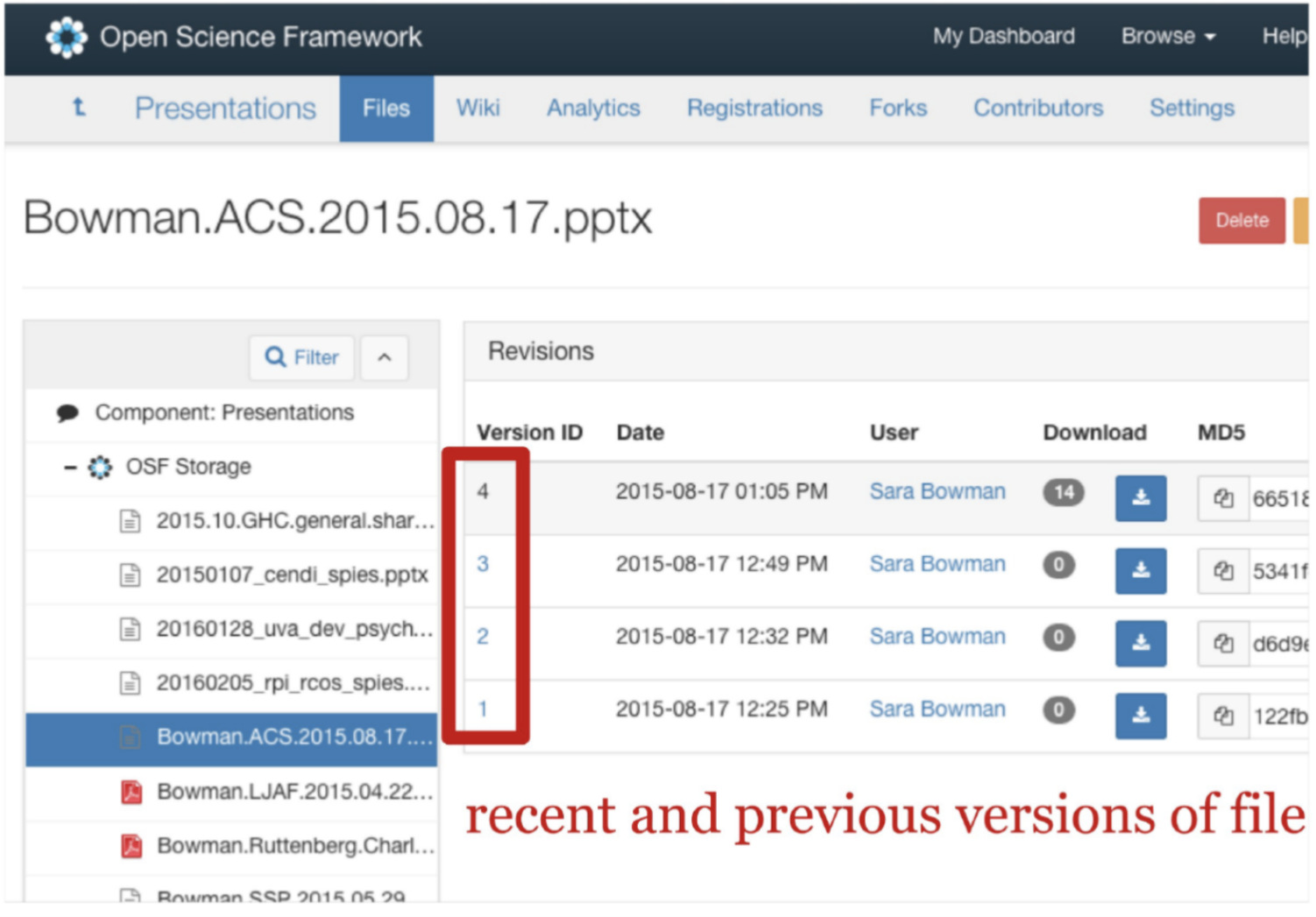
4. Add and document analyses.

5. Share study data, materials, and code.

OSF project landing page

Manage access/permissions

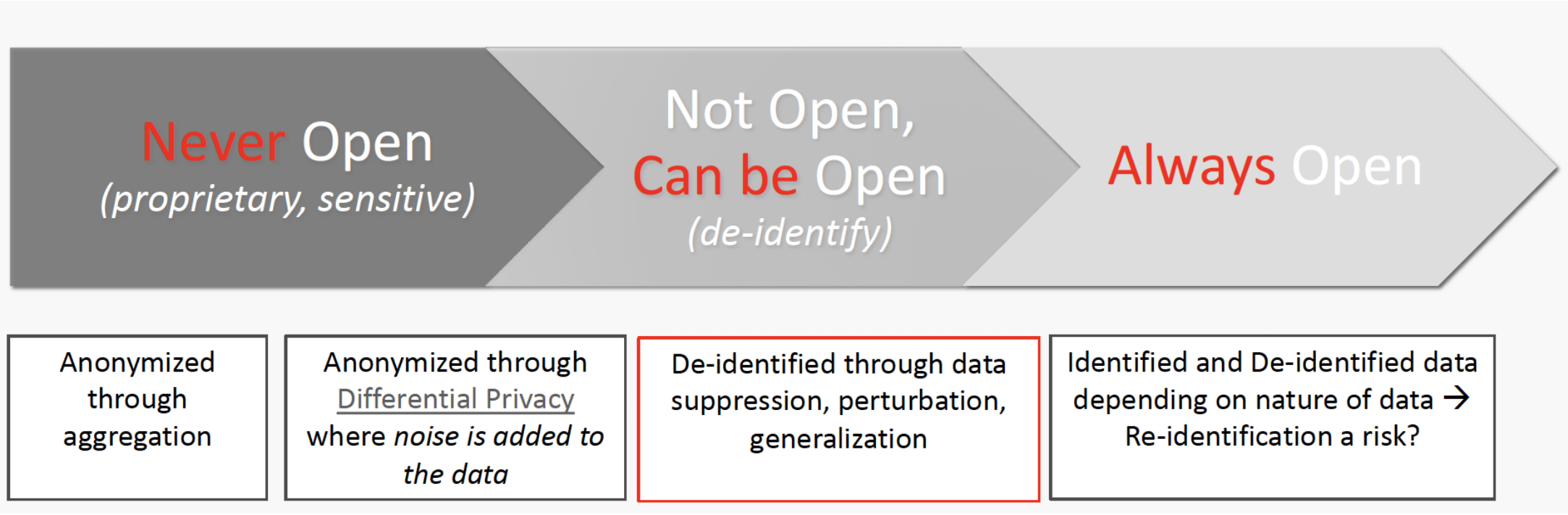
Automate version control



Persistent, unique identifiers for published work

Not everything can (or should) be shared

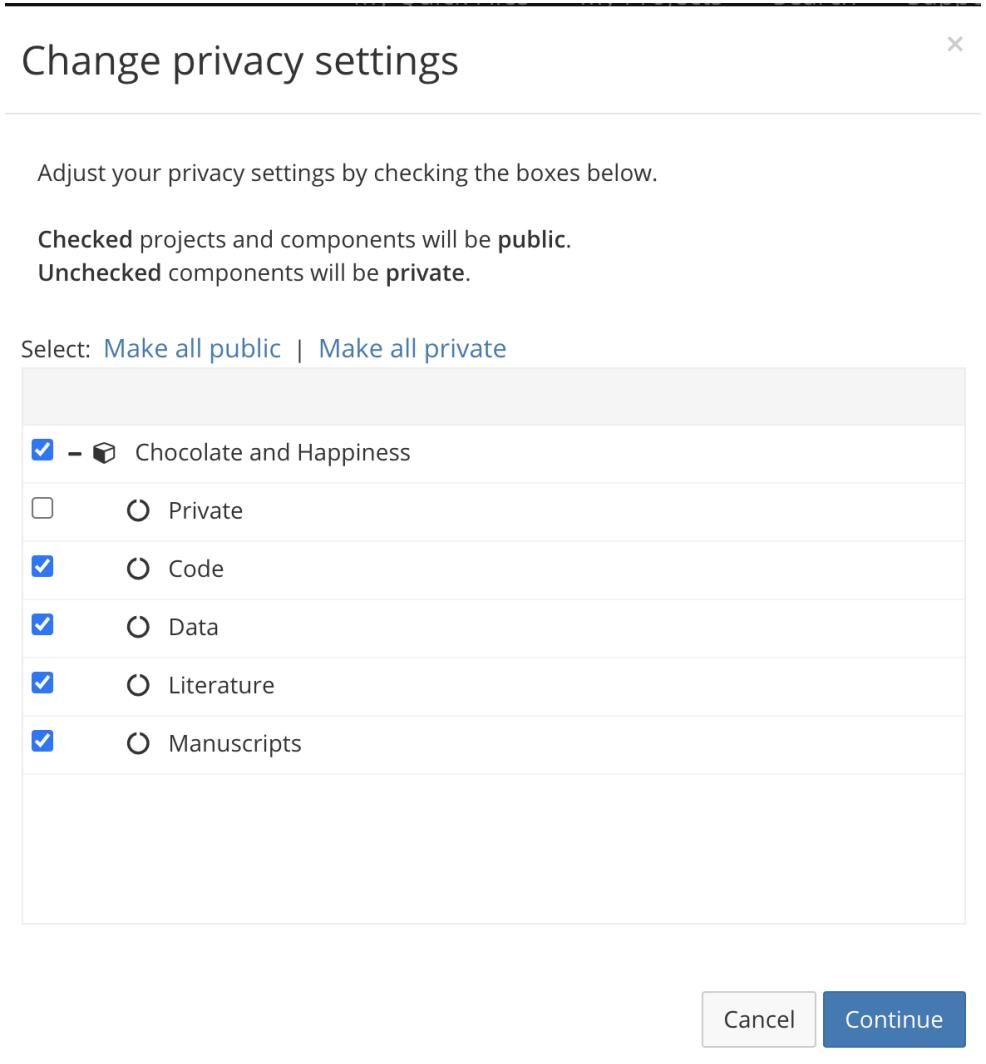
Spectrum for sharing sensitive material:



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

OSF allows both public and private components

Allows for maximum sharing without sacrificing privacy concerns.

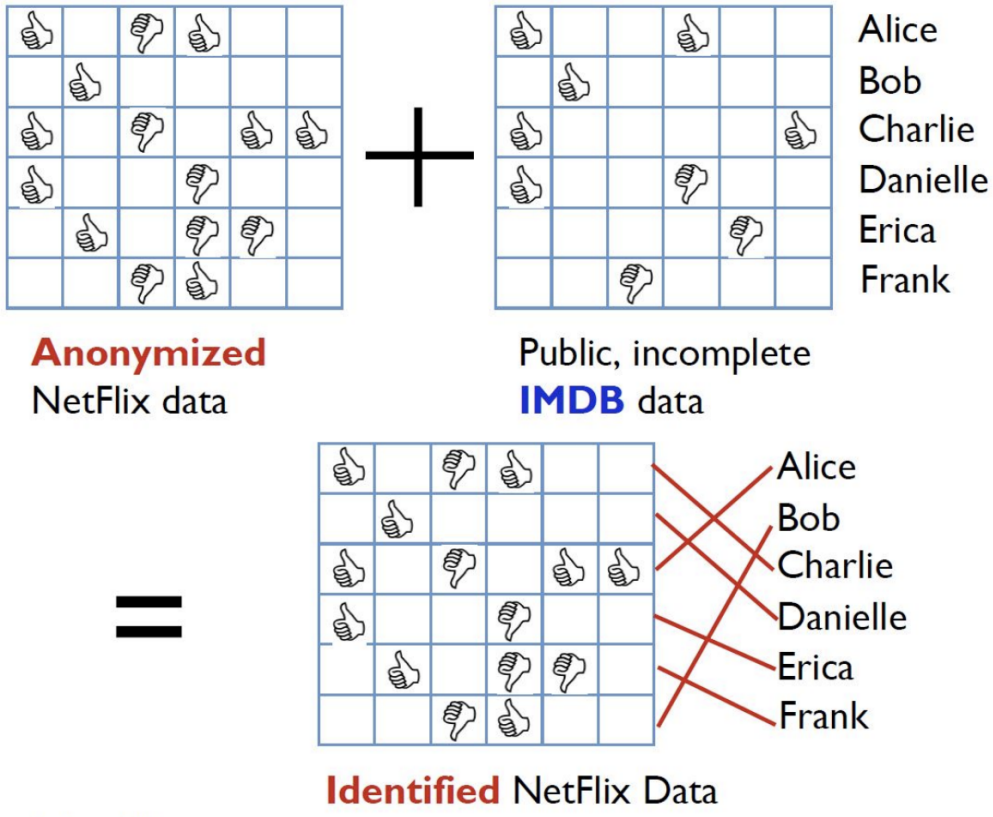


Why we worry (re-identification)

Netflix "contest" data:

- unique "random" ID
- film rated
- date
- ⭐⭐⭐

But people rate in many places, some of which do contain personal data.



Narayanan (2008)

Concerns about privacy

Trade-offs

Can't pretend there aren't social costs to open data.

Dangers of re-identification

Wealth of data after social media revolution.

Limits

Solutions may be too severe to justify utility of sharing.

Synthetic data may also be possible

- mimics an original dataset, preserving its statistical properties and relationships between variables
- classification and regression tree approach
- 0% disclosure risk
- `synthpop` R package 📦
- code and materials on GitHub

Quintana (2020)

Synthetic datasets: A non-technical primer for the biobehavioural sciences to promote reproducibility and hypothesis-generation

Daniel S. Quintana

Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, University of Oslo, and Oslo University Hospital, Oslo, Norway.

Abstract

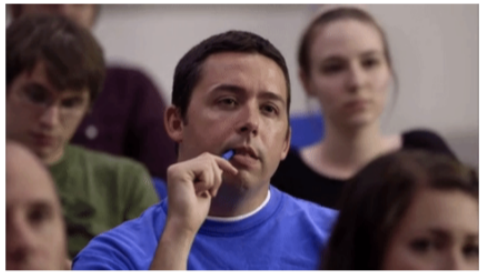
Open research data provides considerable scientific, societal, and economic benefits. However, disclosure risks can sometimes limit the sharing of open data, especially in datasets that include sensitive details or information from individuals with rare disorders. This article introduces the concept of synthetic datasets, which is an emerging method originally developed to permit the sharing of confidential census data. Synthetic datasets mimic real datasets by preserving their statistical properties and the relationships between variables. Importantly, this method also reduces disclosure risk to essentially nil as no record in the synthetic dataset represents a real individual. This practical guide with accompanying R script enables biobehavioural researchers to create synthetic datasets and assess their utility via the *synthpop* R package. By sharing synthetic datasets that mimic original datasets that could not otherwise be made open, researchers can ensure the reproducibility of their results and facilitate data exploration while maintaining participant privacy.

5. Reproducible Example

Today's Research Project

Does chocolate increase graduate student happiness?

Subjects



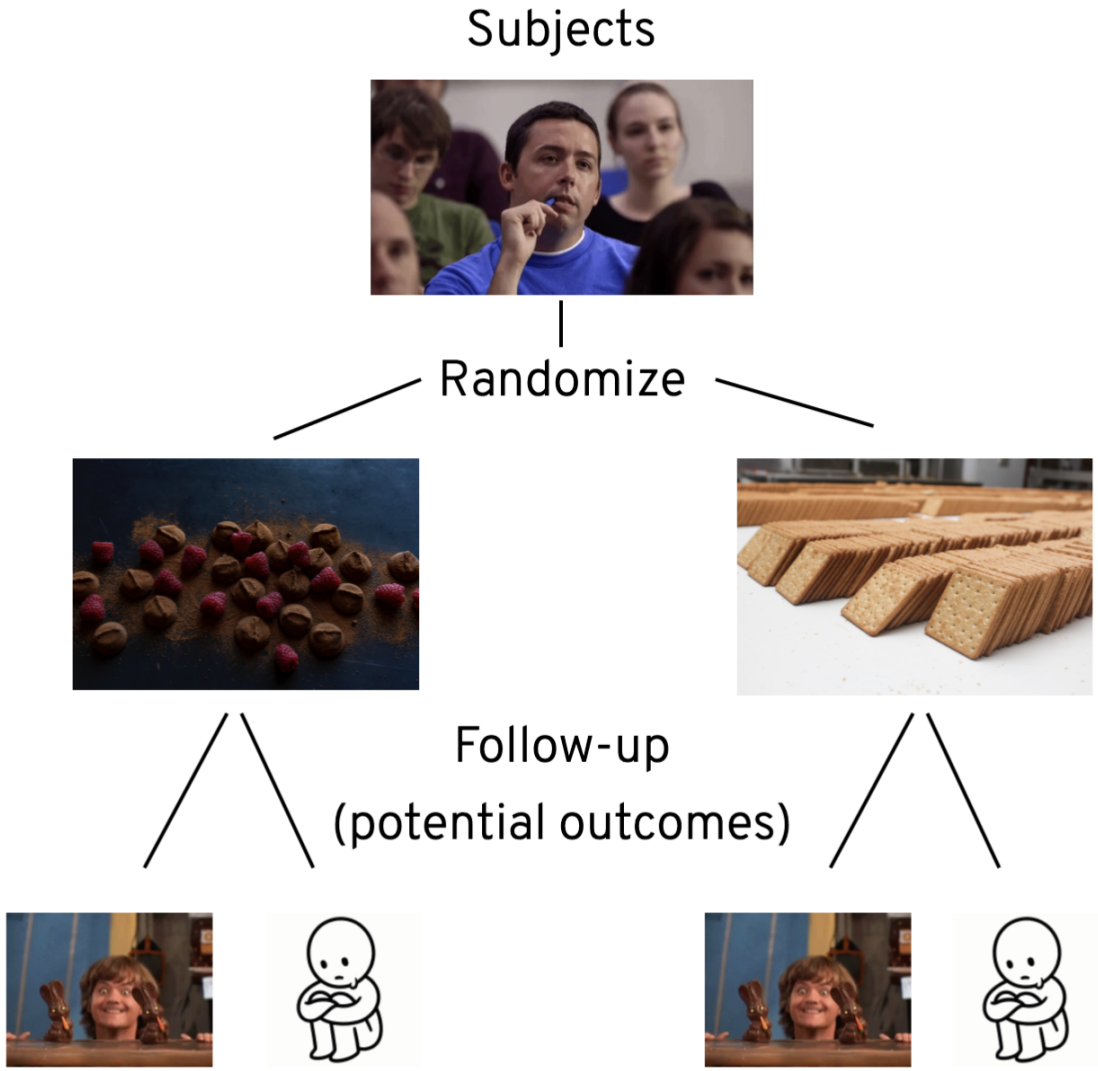
Randomize

Follow-up

(potential outcomes)

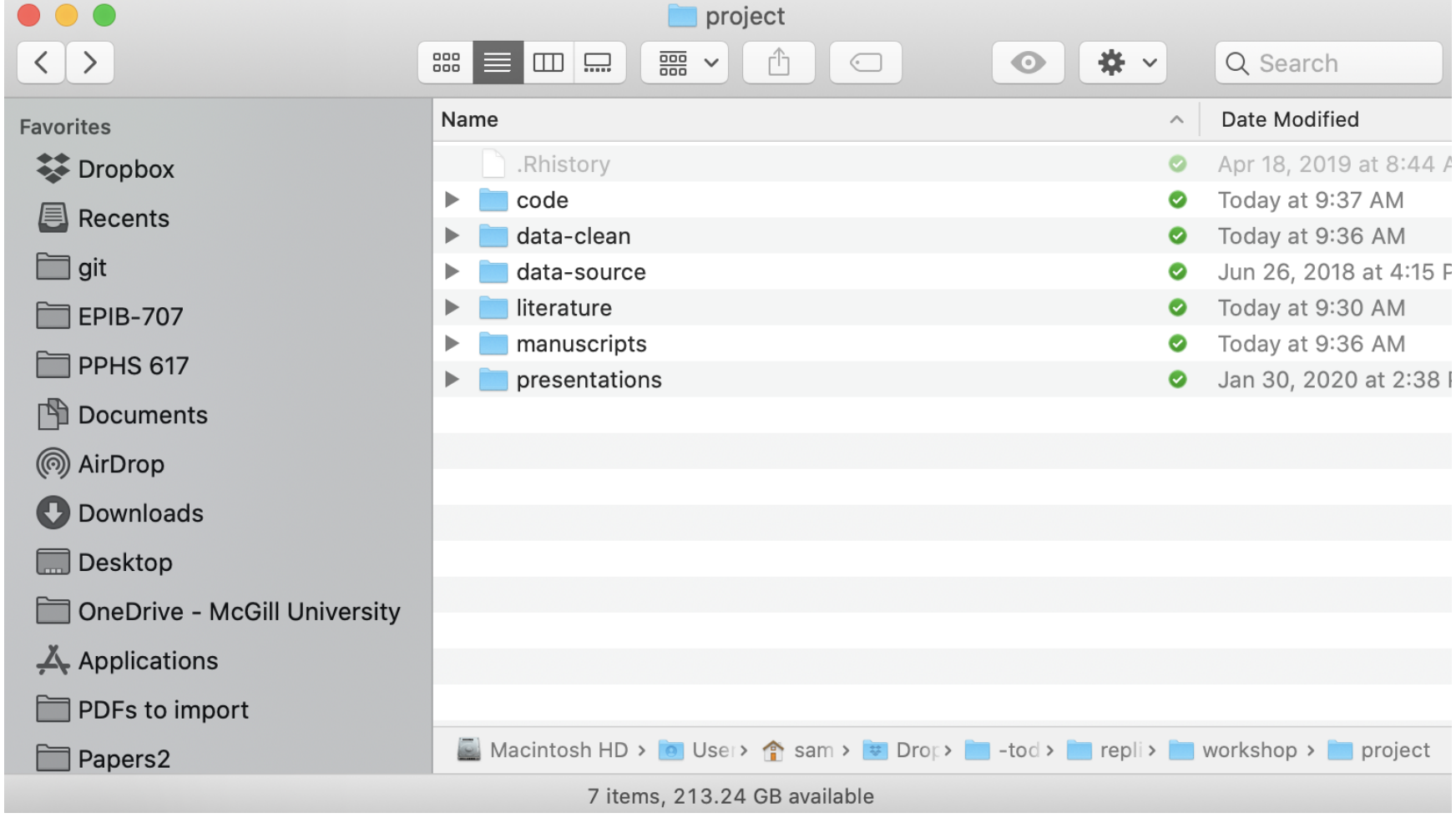
Example

- Folder structure
- Organizing via OSF
- Cleaning
- Analysis
  - Descriptives
  - Tables
  - Figures
- Write-up



Folder structure

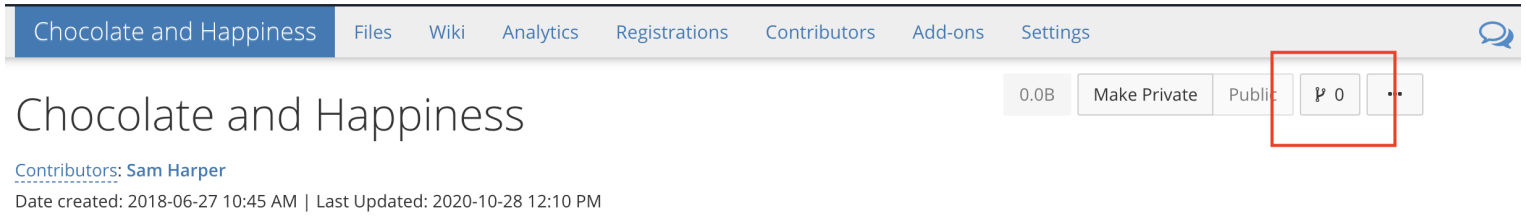
Separate code from data and manuscript:



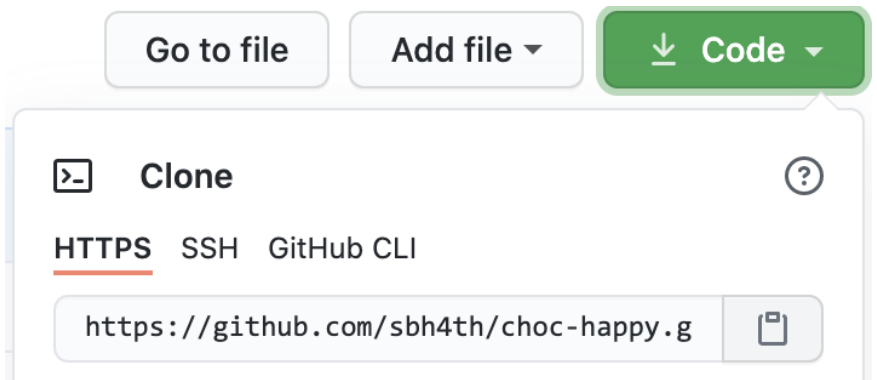
Live Demo

Follow along

Create a copy of the OSF repository (called 'forking')

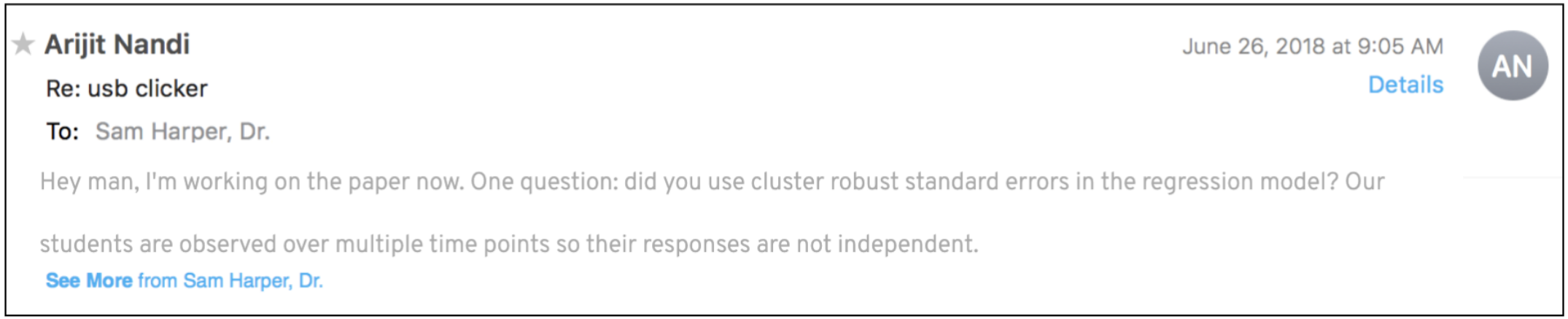


Clone the GitHub repository (download to your machine)



Final Thoughts

Working with "friends"



1. Ugh.
2. He's right.
3. Need to regenerate the model and table results.

- Do you know where your materials are?
- Is it 6 months later?
- Could someone else take over and figure out what you did?

Other resources for reproducibility in R

Version Control

- Happy Git and GitHub for the useR

RMarkdown

- R Markdown: The Definitive guide
- RMarkdown Driven Development (RmdDD): Blog post by Emily Riederer

R Packages

- R packages by Hadley Wickham and Jenny Bryan

Research Compendia

- Karthik Ram: *rstudio::conf 2019 talk*

Docker & Binder

- Getting started with binder docs
- rOpenSci Docker tutorial

Tutorials

- Rstudio Essentials Webinar series
- rrresearch: ACCE DTP course on Research Data & Project Management

Source: Anna Krystalli

This seems like a lot of extra work.

What's in it for me?

Selfish reasons to work reproducibly:

- Helps to avoid disaster.
- Makes it easier to write papers.
- Can help reviewers
- Enables continuity of your work.
- Builds your scholarly reputation.

Markowetz (2015)

Take home messages

Reproducible ≠ rigid

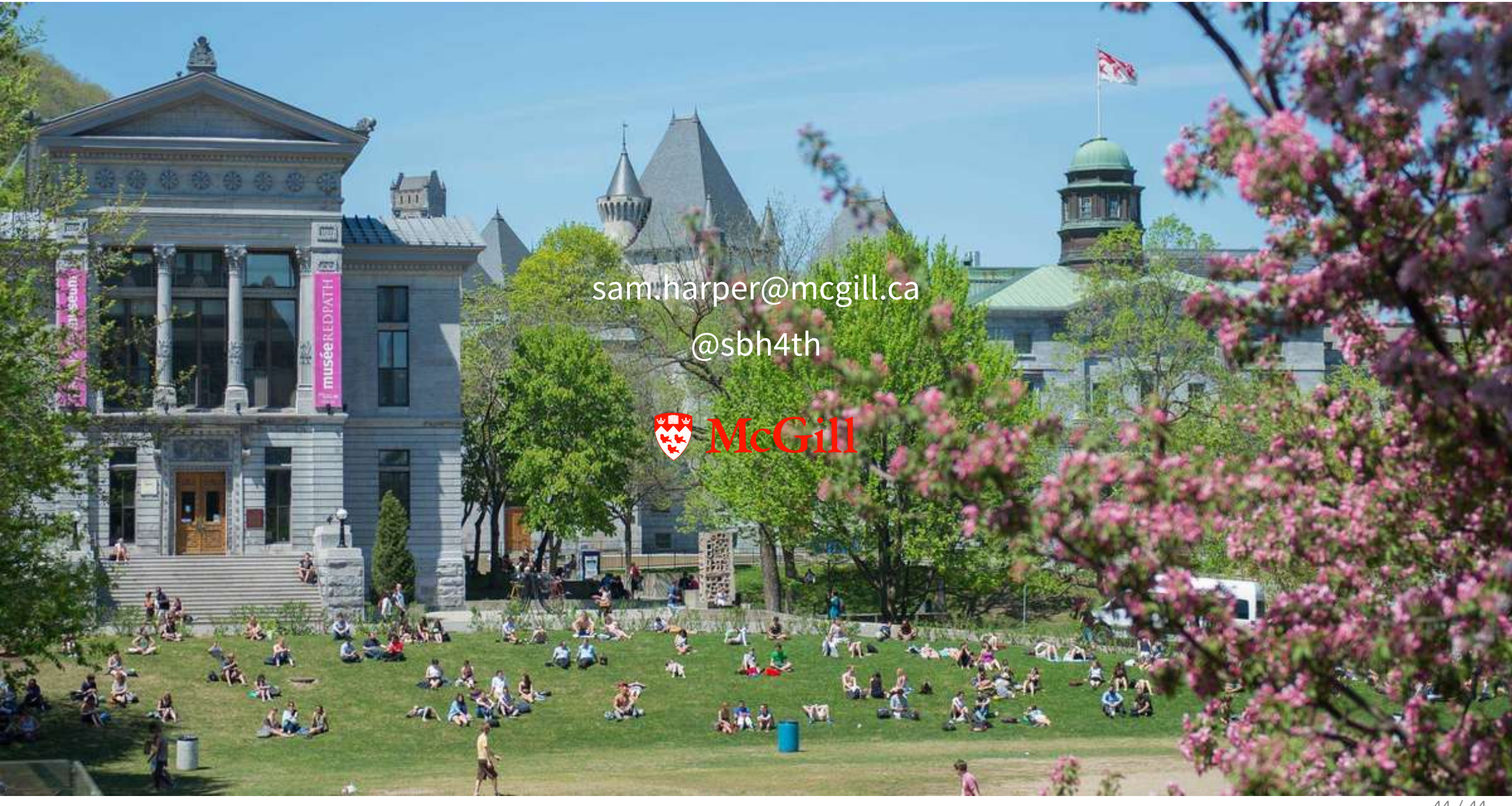
Don't like *R*? Need to use SAS? There is no single set of tools--more important to find a workflow that works for **you** and your team.

Security still matters

Some things can't be shared. Providing a rationale for what is/not shared means being transparent.

Lead by example

Transparency and openness are positive norms. We need help.



sam.harper@mcgill.ca

@sbh4th

McGill