

Calibrating the scientific ecosystem through meta-research

Tom Hardwicke

Meta-Research Innovation Center Berlin

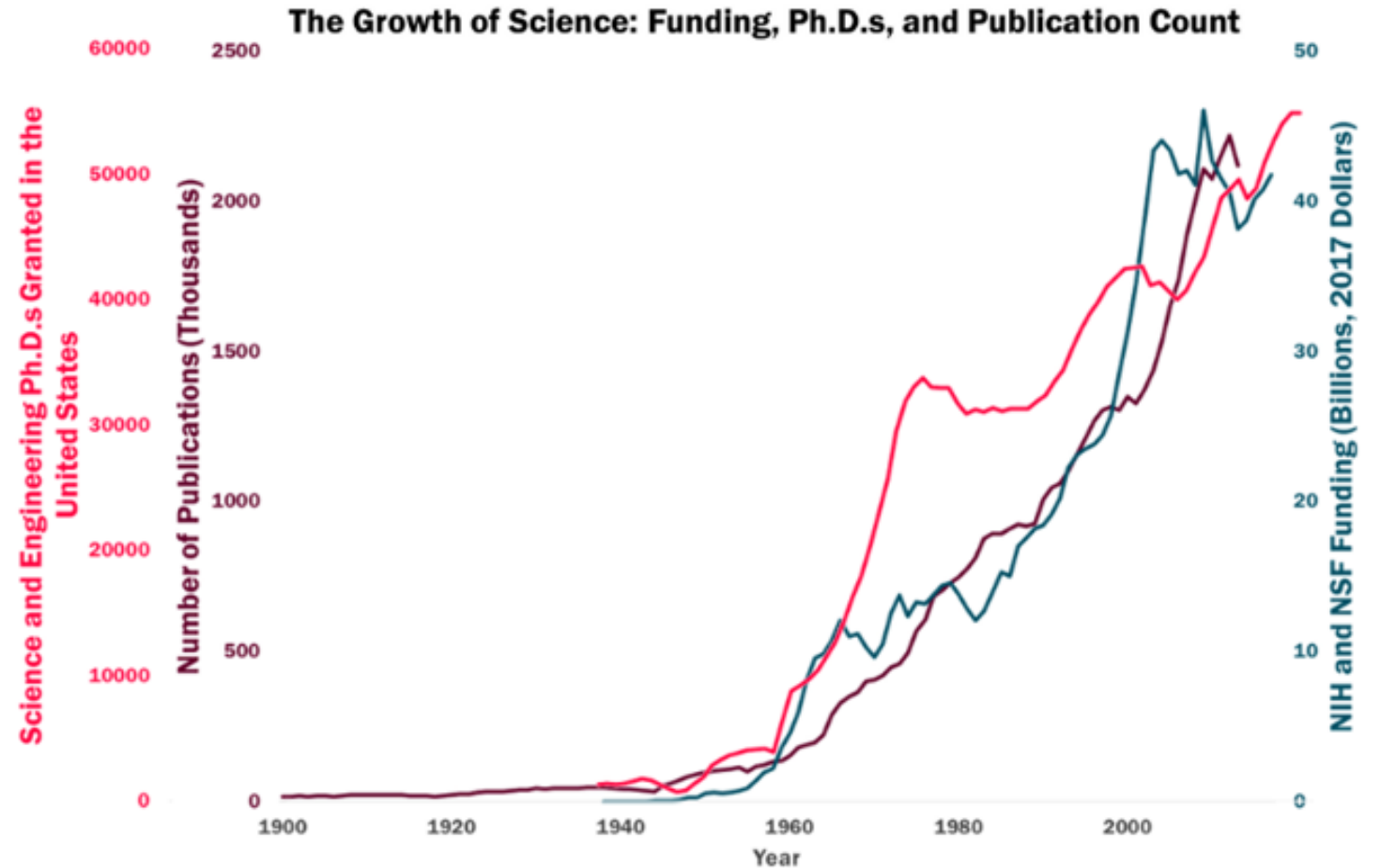
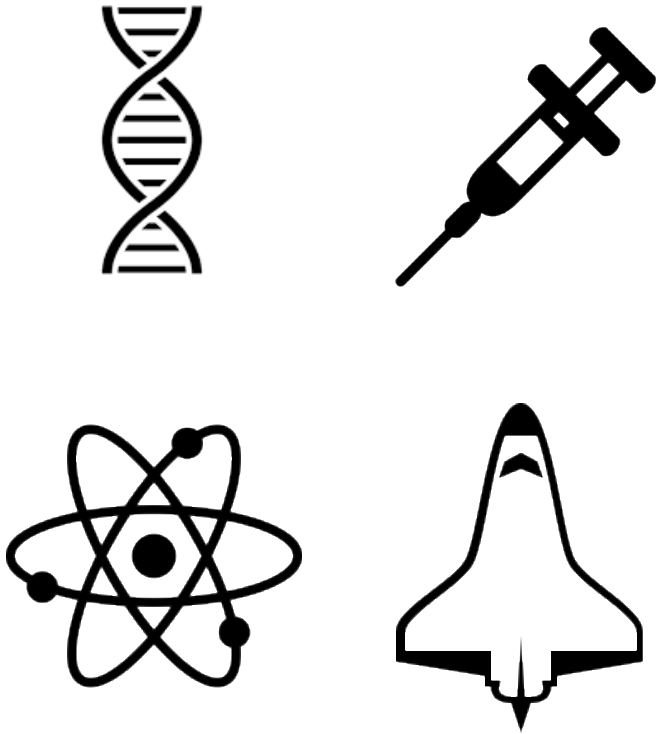
QUEST Center for Transforming Biomedical Research

Berlin Institute of Health

Charité – Universitätsmedizin Berlin



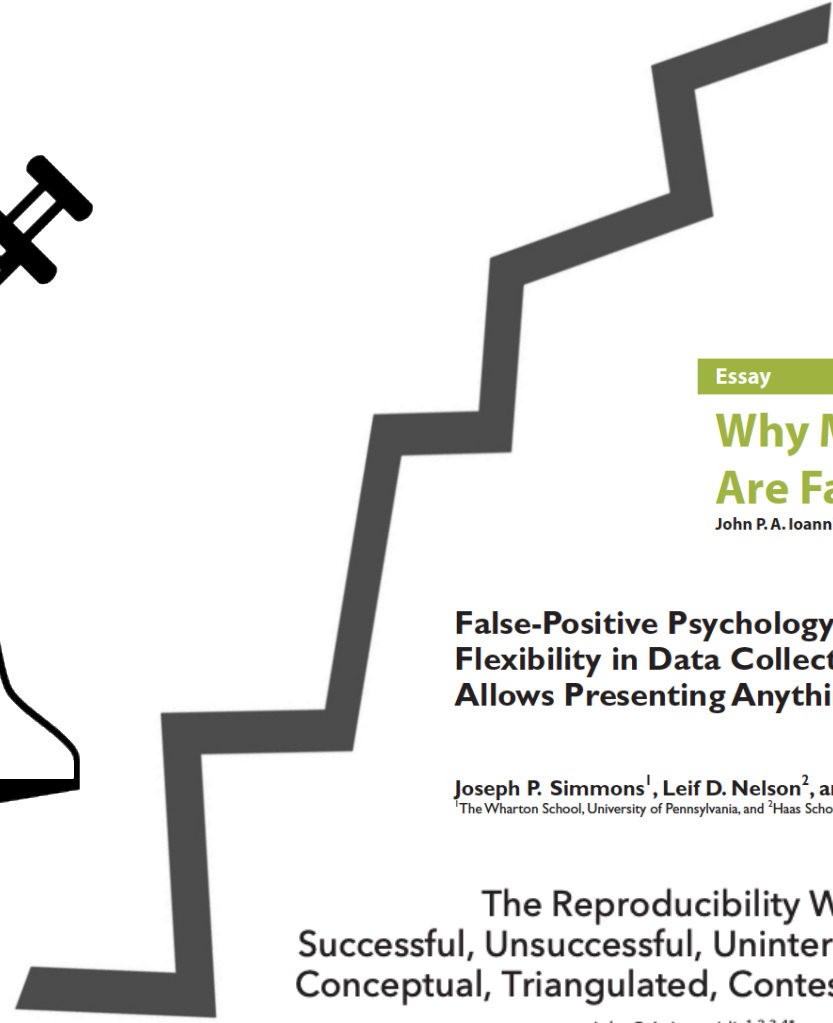
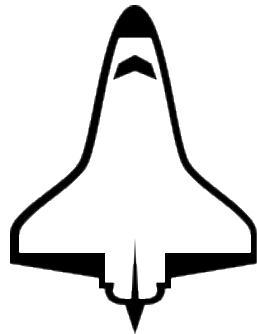
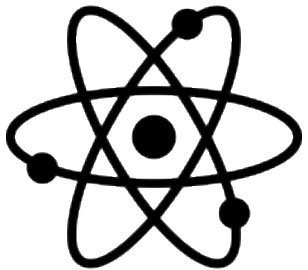
Science: A success story?



Collison & Nielsen (2018)



Science: A success story?



The Statistical Crisis in Science

Data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don’t hold up.

Andrew Gelman and Eric Loken

Let’s Take the Con out of Econometrics

By EDWARD E. LEAMER*

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

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

John P.A. Ioannidis^{1,2,3,4*}

IS THERE A REPRODUCIBILITY CRISIS?



©nature

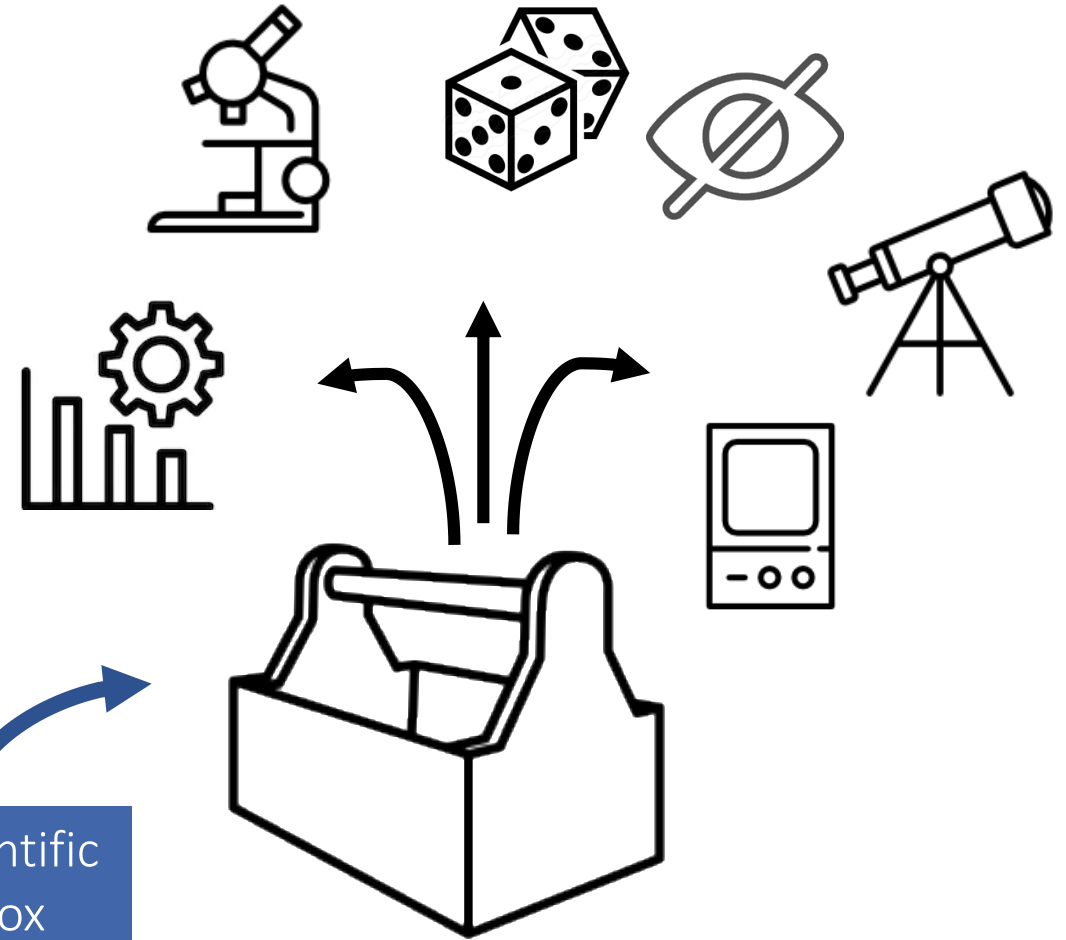


Science as a human endeavour



Only human

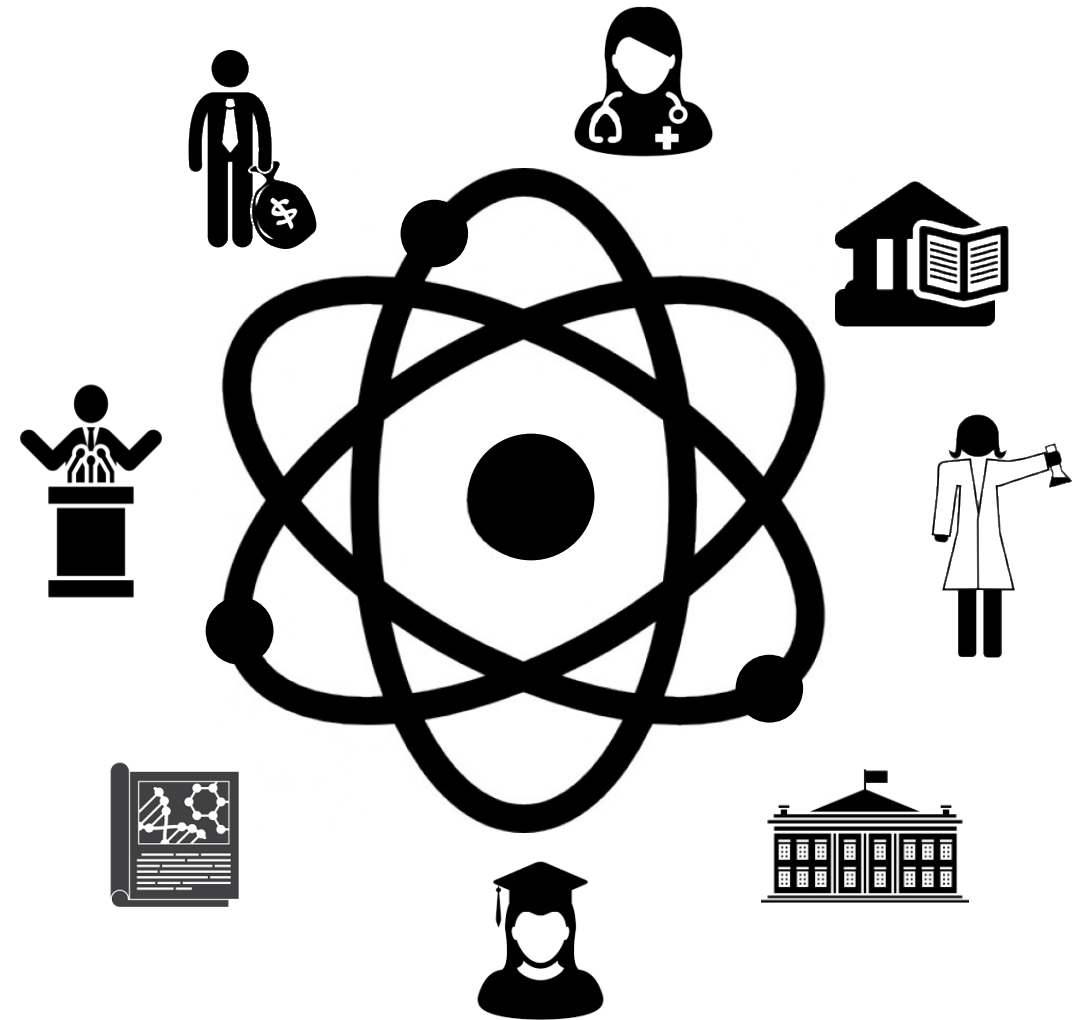
The scientific
toolbox



The scientific ecosystem

Stakeholders

Research departments
Academic societies
Universities
Publishers
Librarians
Journals
Funders
Scientists
Media
Students
Politicians
Practitioners
General public



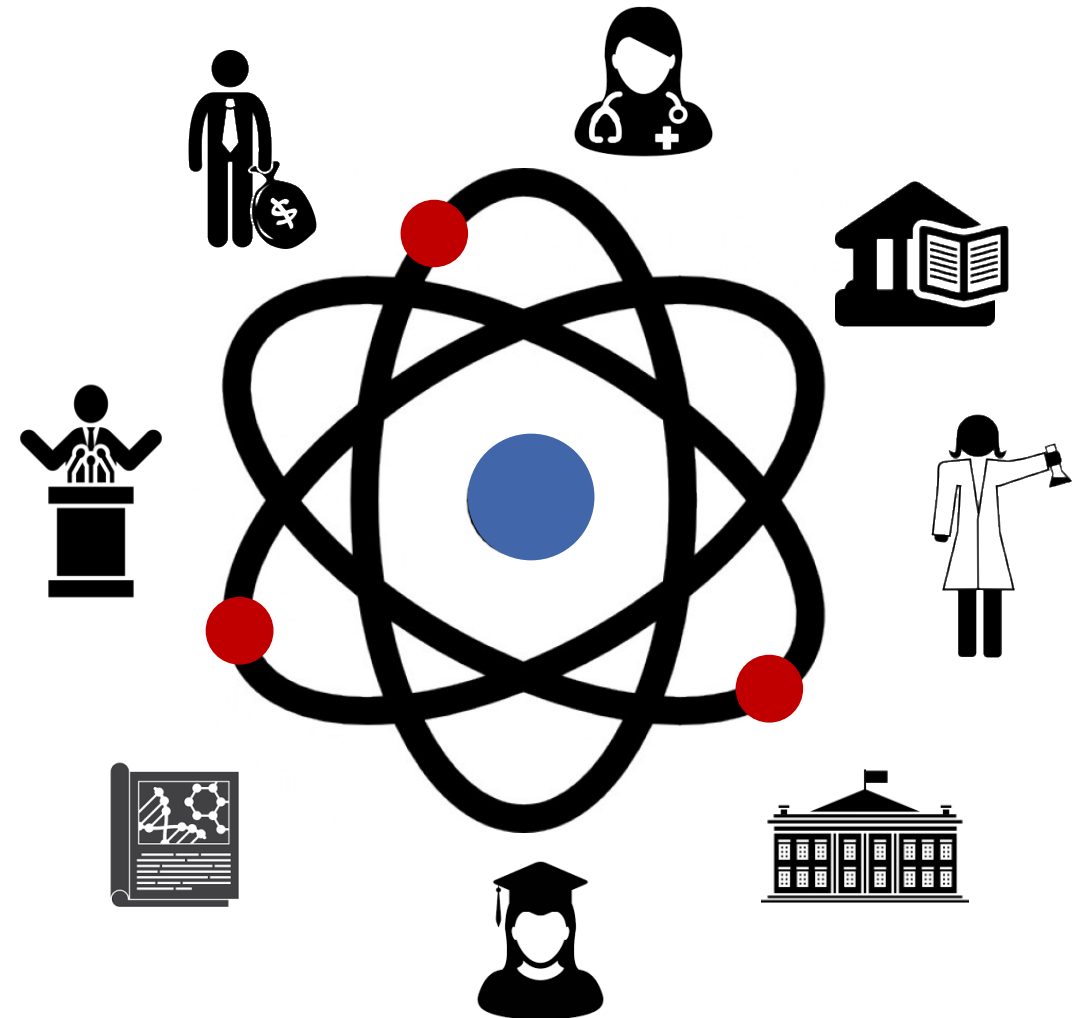
Tragedy of the scientific commons

Aesthetics over authenticity

'novel'
'positive'
'clean'

Vs.

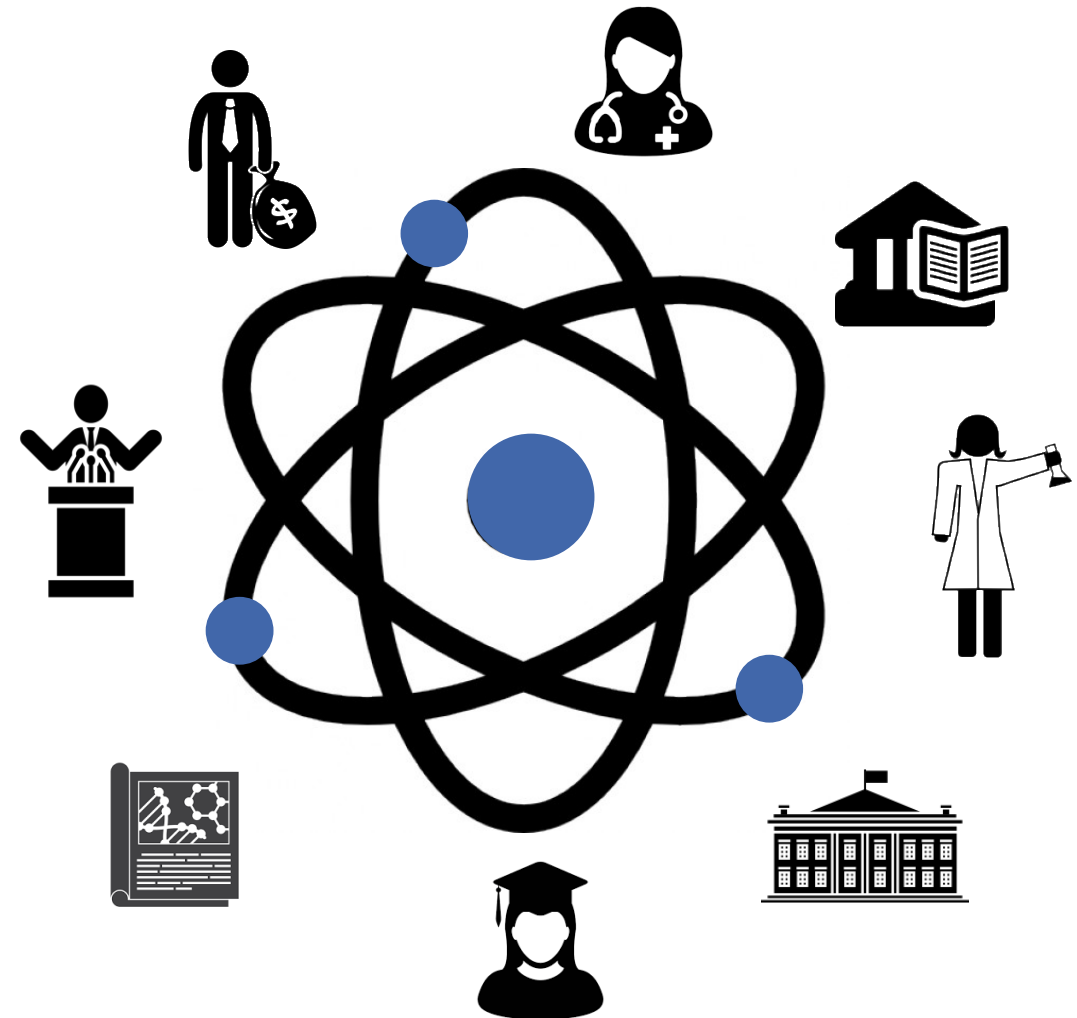
'incremental'
'negative'
'messy'



Calibrating the scientific ecosystem

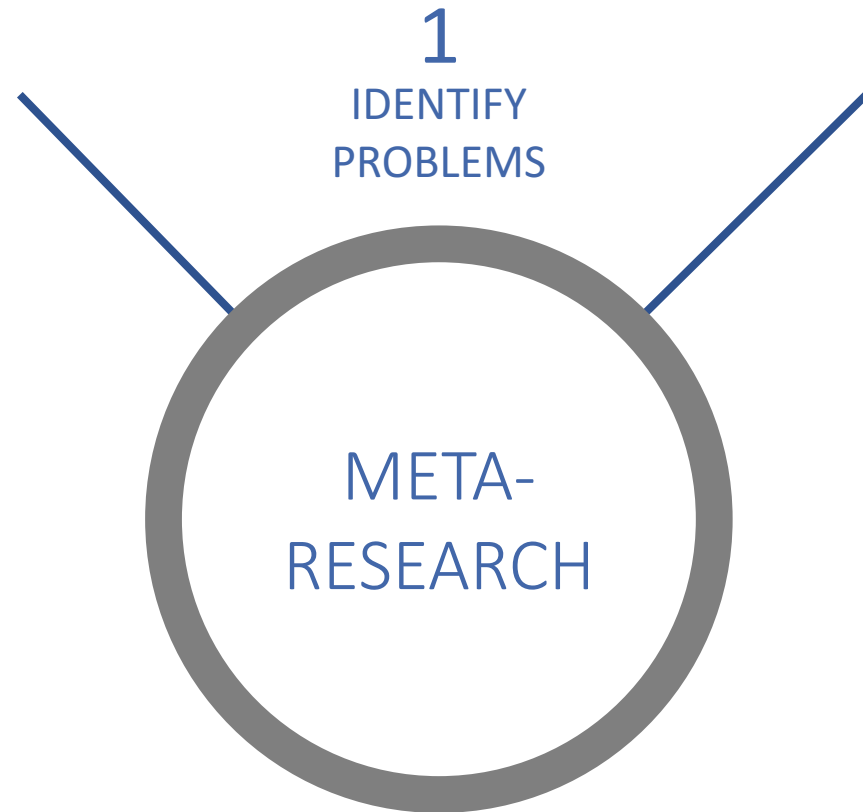
Authenticity over aesthetics

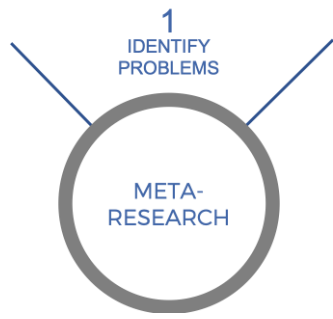
Transparency and Openness Promotion (TOP) Guidelines
Peer Reviewers' Openness (PRO) Initiative
Reporting Guidelines (see EQUATOR)
Statistical review
Pre-registration
Registered Reports
Badges
Redefine / replace statistical significance
Results-blind peer review
Collaboration
Replication
Disclosure statements
Improved design (e.g., statistical power, randomization)
Blind-analysis



A translational framework for meta-research

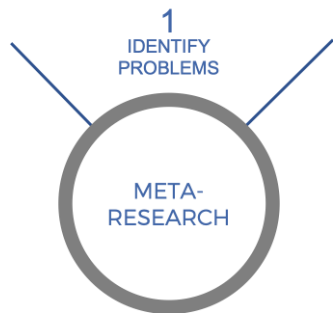




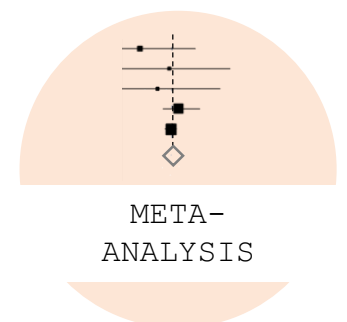
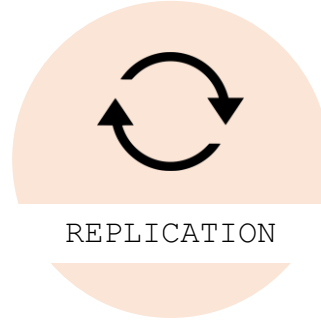
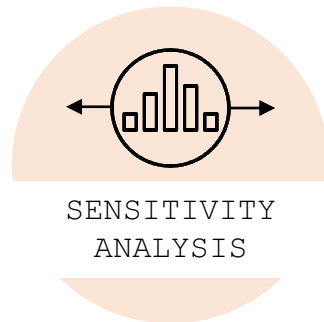


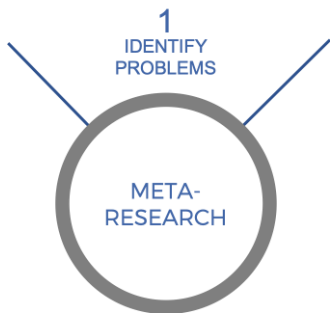
The self-correction machinery of science cannot operate effectively without transparency.



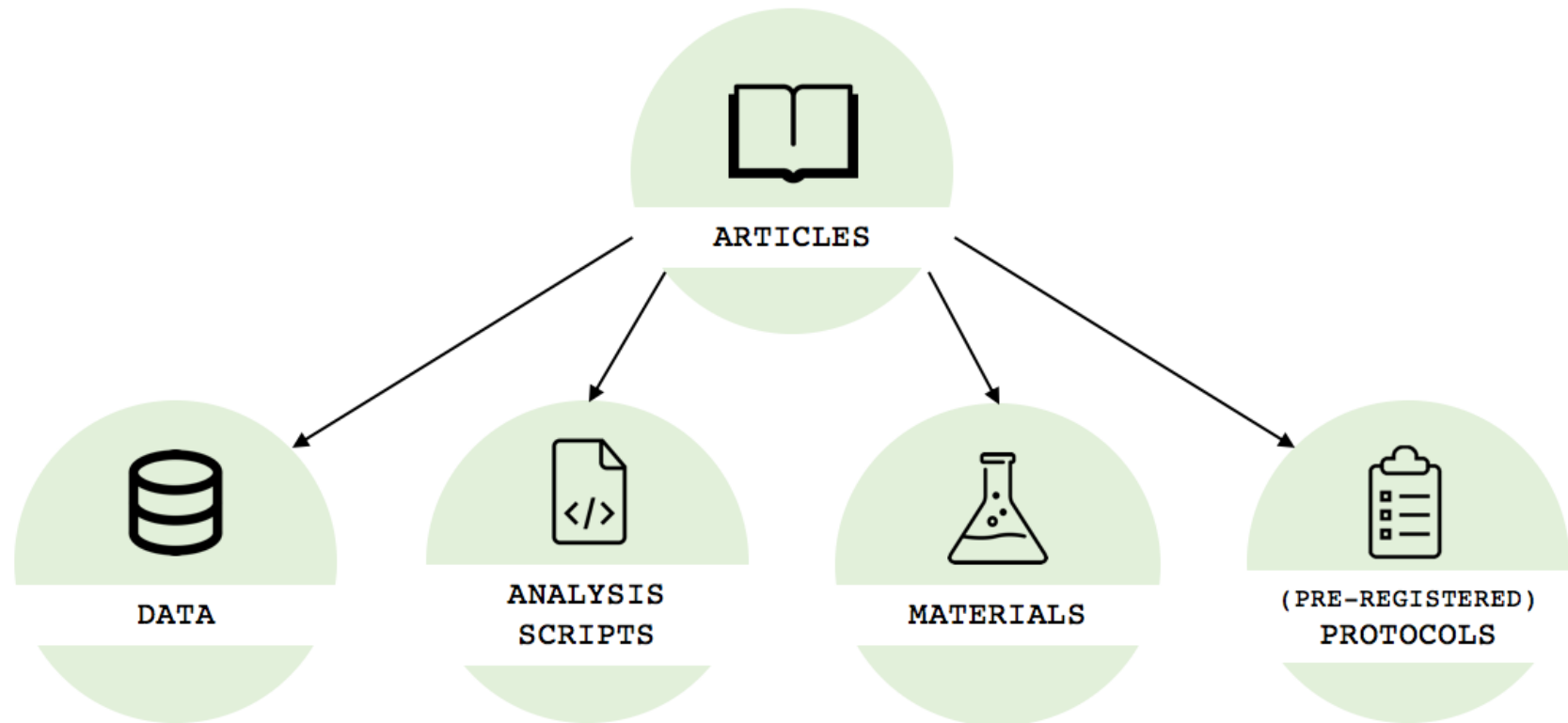


Self-correction machinery



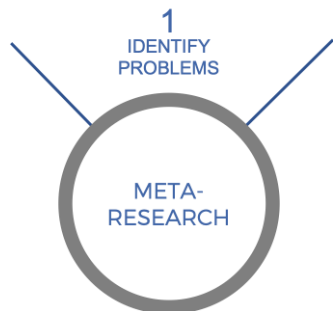


The fuel of scientific self-correction



Research resources ->





Low transparency undermines self-correction

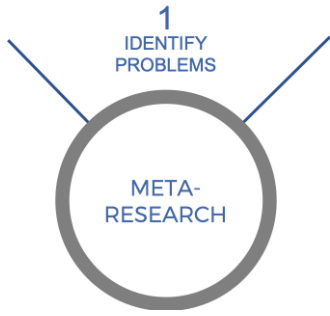
Reproducibility Project: Cancer Biology

Investigating reproducibility in preclinical cancer research.

Plan to replicate 50 high-impact cancer papers shrinks to just 18

By [Jocelyn Kaiser](#) | Jul. 31, 2018 , 5:45 PM





Transparency enables self-correction

Growth in a Time of Debt

By CARMEN M. REINHART AND KENNETH S. ROGOFF 

Original authors' claim: For government debt in excess of 90%, annual growth was 'roughly cut in half'.

Cited by many leading politicians in justification of austerity measures.

Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff

Thomas Herndon, Michael Ash and Robert Pollin[★]

"While using RR's working spreadsheet, we identified coding errors, selective exclusion of available data, and unconventional weighting of summary statistics."







The attrition of the modern scholarly record

Most published scientific papers are not accompanied by raw data

Study	Field	Papers checked	% available
Alsheikh-Ali et al. (2011)	Multiple	500	9%
Iqbal et al. (2014)	Biomedicine	441	0%

The attrition of the modern scholarly record

Data often not available on request

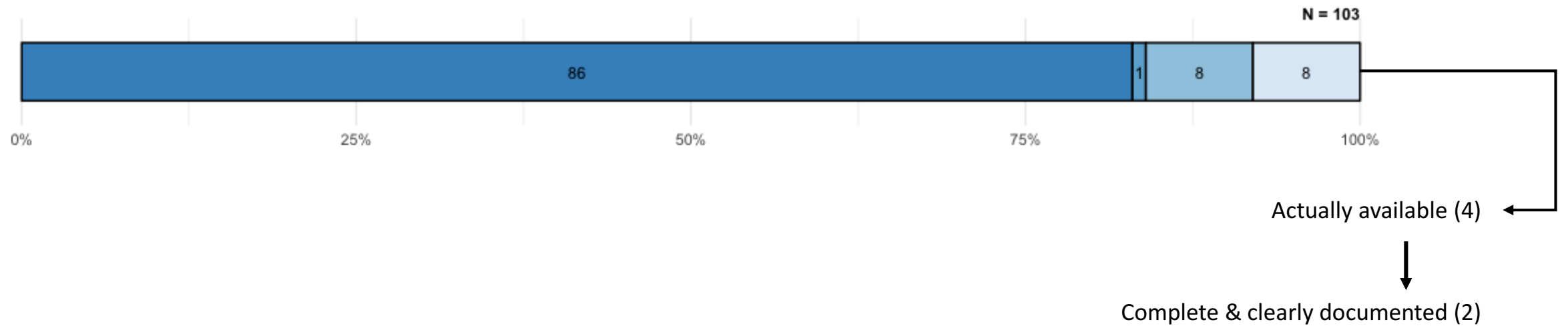
Study	Field	Papers checked	% data available*
Vanpaemel et al. (2015)	Psychology	394	38%
Vines et al. (2014)	Ecology	516	19%
Krawczyk et al. (2012)	Economics	200	44%
Hardwicke & Ioannidis (2018)	Psychology & Psychiatry	111	14%

*unrestricted

Transparent research practices in the social sciences (2014-2017)

Data availability

Statement reports: ■ No statement ■ Not available ■ External data source ■ Available

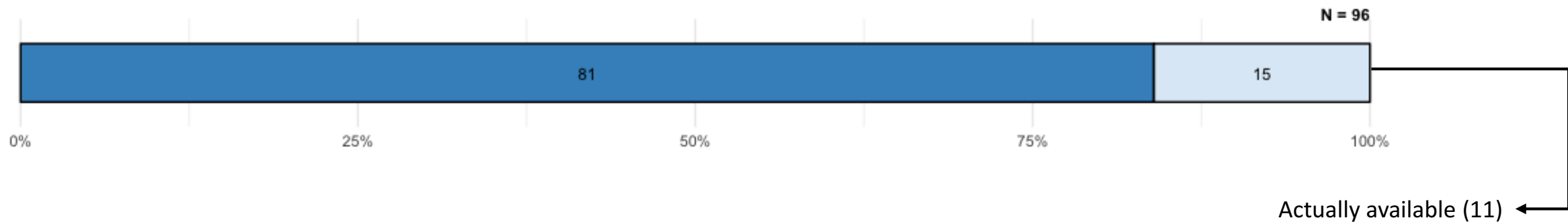


Hardwicke et al. (submitted)

Transparent research practices in the social sciences (2014-2017)

Materials availability

Statement reports: ■ No statement ■ Not available ■ Available

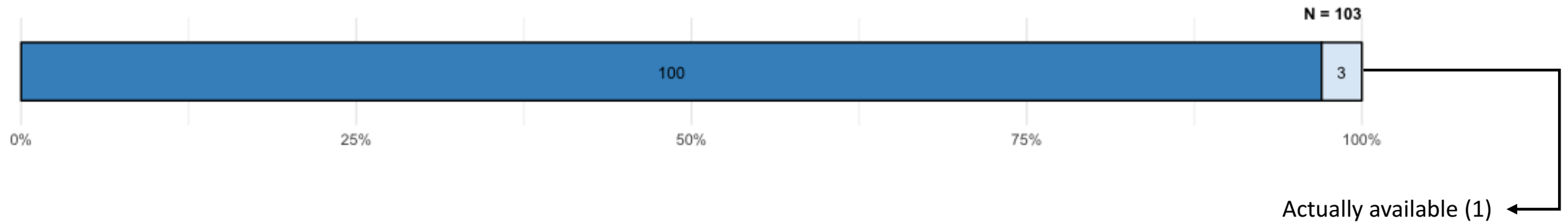


Hardwicke et al. (submitted)

Transparent research practices in the social sciences (2014-2017)

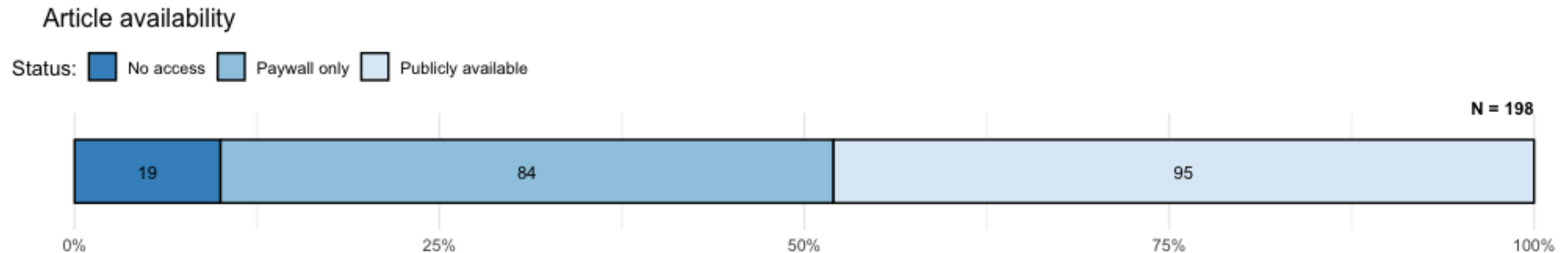
Analysis script availability

Statement reports: ■ No statement ■ Not available ■ Available



Hardwicke et al. (submitted)

Transparent research practices in the social sciences (2014-2017)

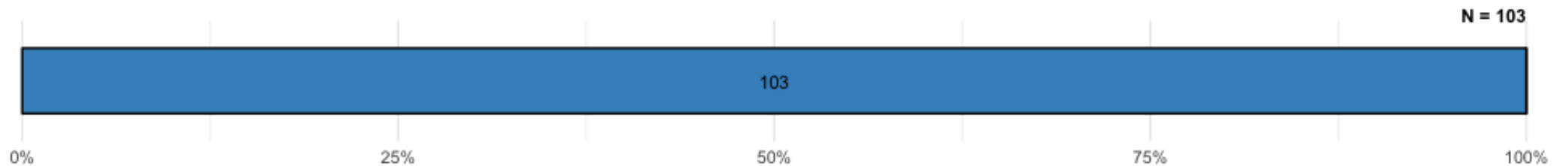


Hardwicke et al. (submitted)

Transparent research practices in the social sciences (2014-2017)

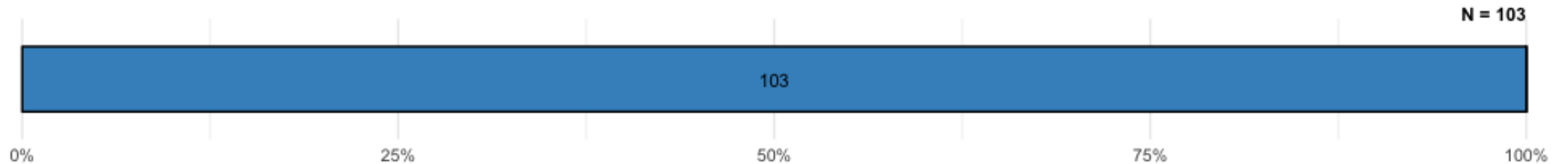
Protocol availability

Statement reports: ■ No statement ■ Available



Pre-registration

Statement reports: ■ No statement ■ Not pre-registered ■ Pre-registered



Hardwicke et al. (submitted)

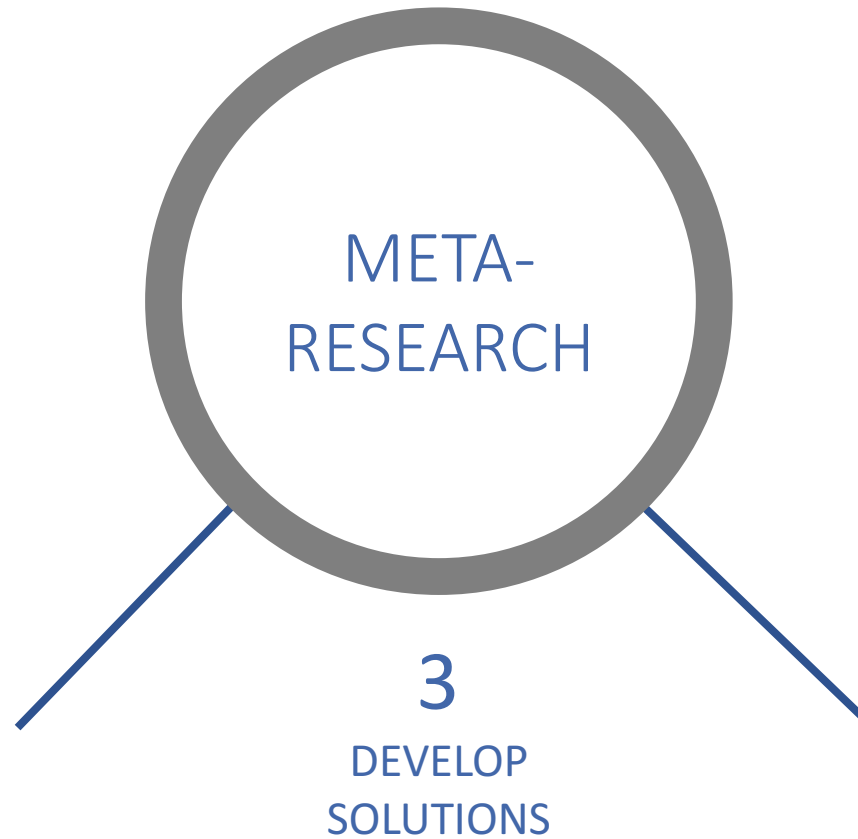


Acknowledging negative constraints

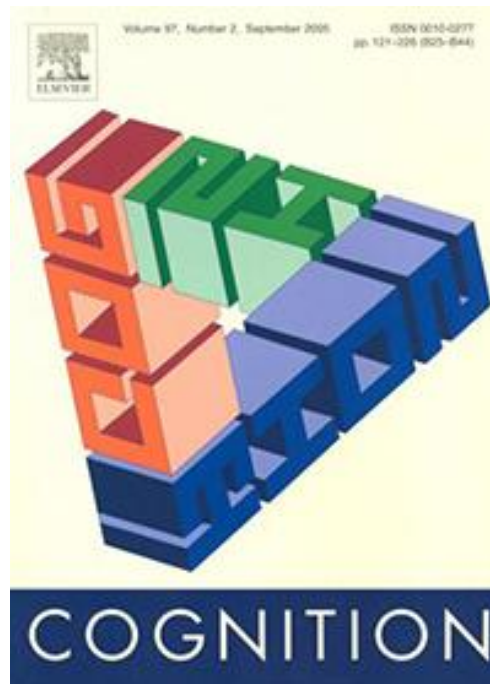
- Sometimes research resource cannot be made available due to overriding legal, ethical, or practical issues (negative constraints).
- The “transparency as default” maxim suggests that when negative constraints prevent sharing, a minimum action is to explicitly disclose this in any associated manuscripts (see Morey et al., 2016, PRO-I; Nosek et al., 2015, TOP).







Journal open data policies



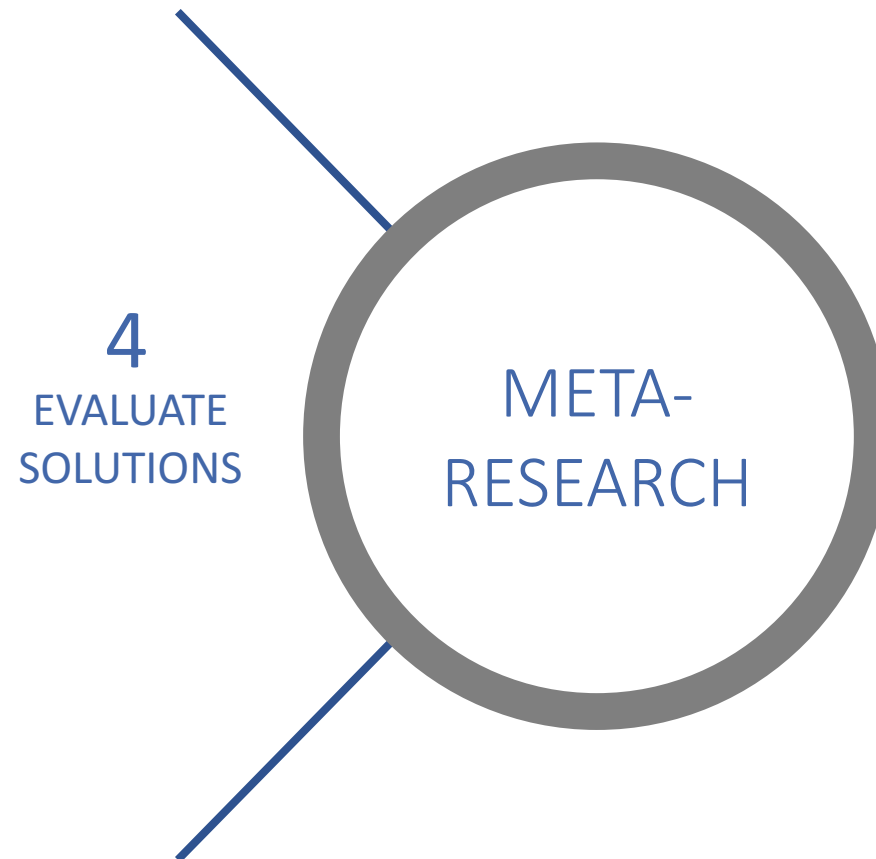
A mandatory open data policy was introduced at the journal Cognition on 1st March, 2015:

“All empirical papers must archive their data upon acceptance in order to be published unless the authors provide a compelling reason why they cannot.”

Additionally:

“The data must be in a form that allows all reported statistical analyses to be reproduced...”





Research article

Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*

Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, Gustav Nilsson, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J. Yoon, Michael Henry Tessler, Richie L. Lenne, Sara Altman, Bria Long, and Michael C. Frank



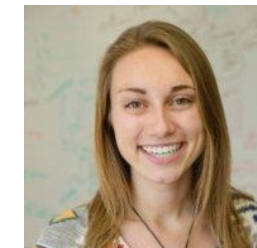
Maya Mathur



Kyle MacDonald



Gustav Nilsson



Mallory Kidwell



George Banks



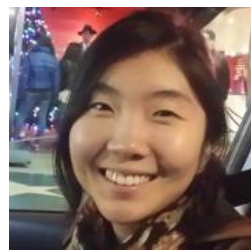
Michael Henry Tessler



Richie Lenne



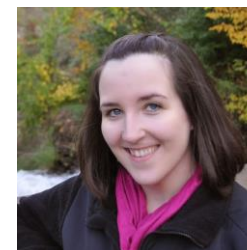
Sara Altman



Erica Yoon



Elizabeth Clayton



Alicia Hofelich Mohr

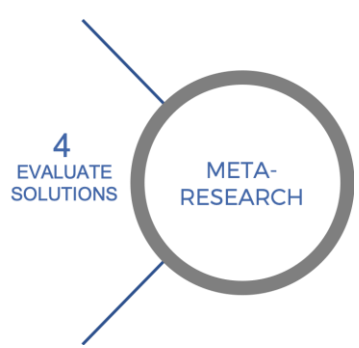


Bria Long

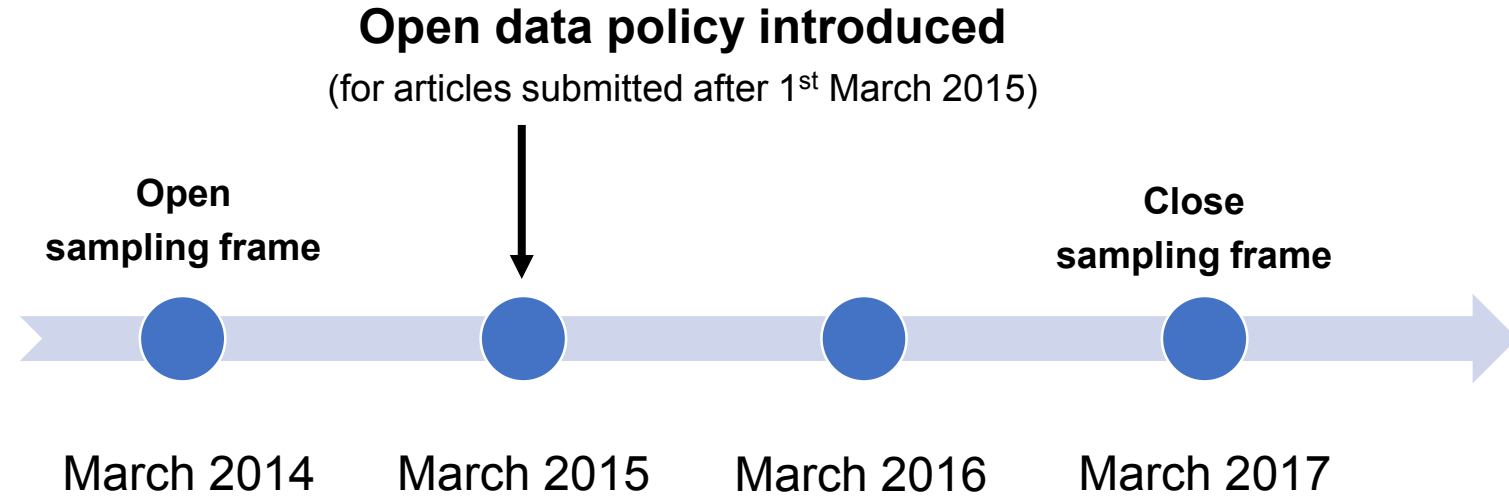


Michael Frank

Hardwicke et al. (2018)



Sampling frame



Sample: 591 articles published between March 2014 and March 2017 (417 submitted during pre-policy period, 174 submitted during post-policy period).

Hardwicke et al. (2018)





Methods

Coders manually extracted information on:

- **Availability**

(was there a data availability statement?)

- **Accessibility**

(could we successfully download and open the data file?)

- **Completeness**

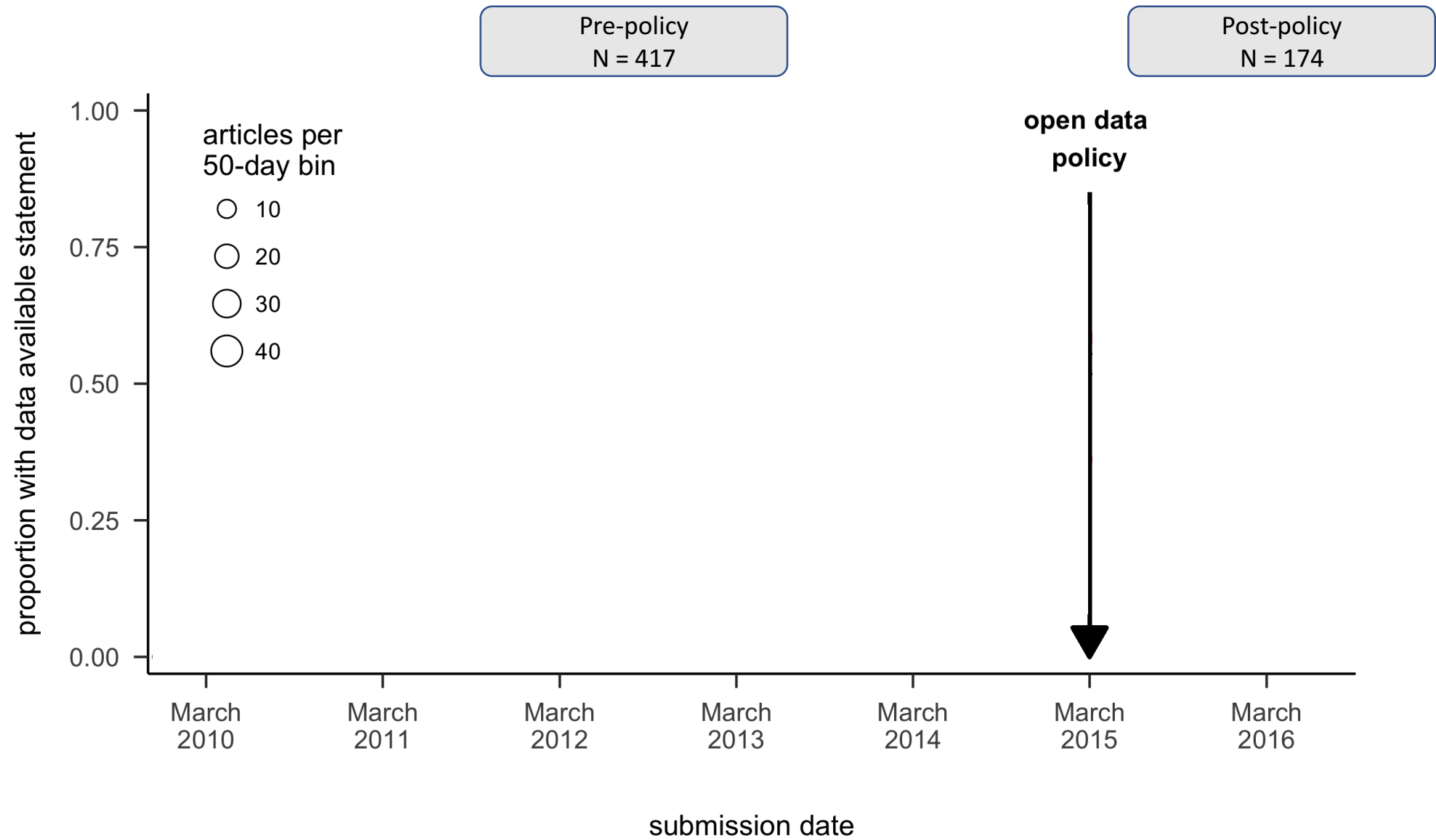
(is raw data provided for all variables measured during the study?)

- **Understandability**

(are the data files sufficiently labelled/documentated?)

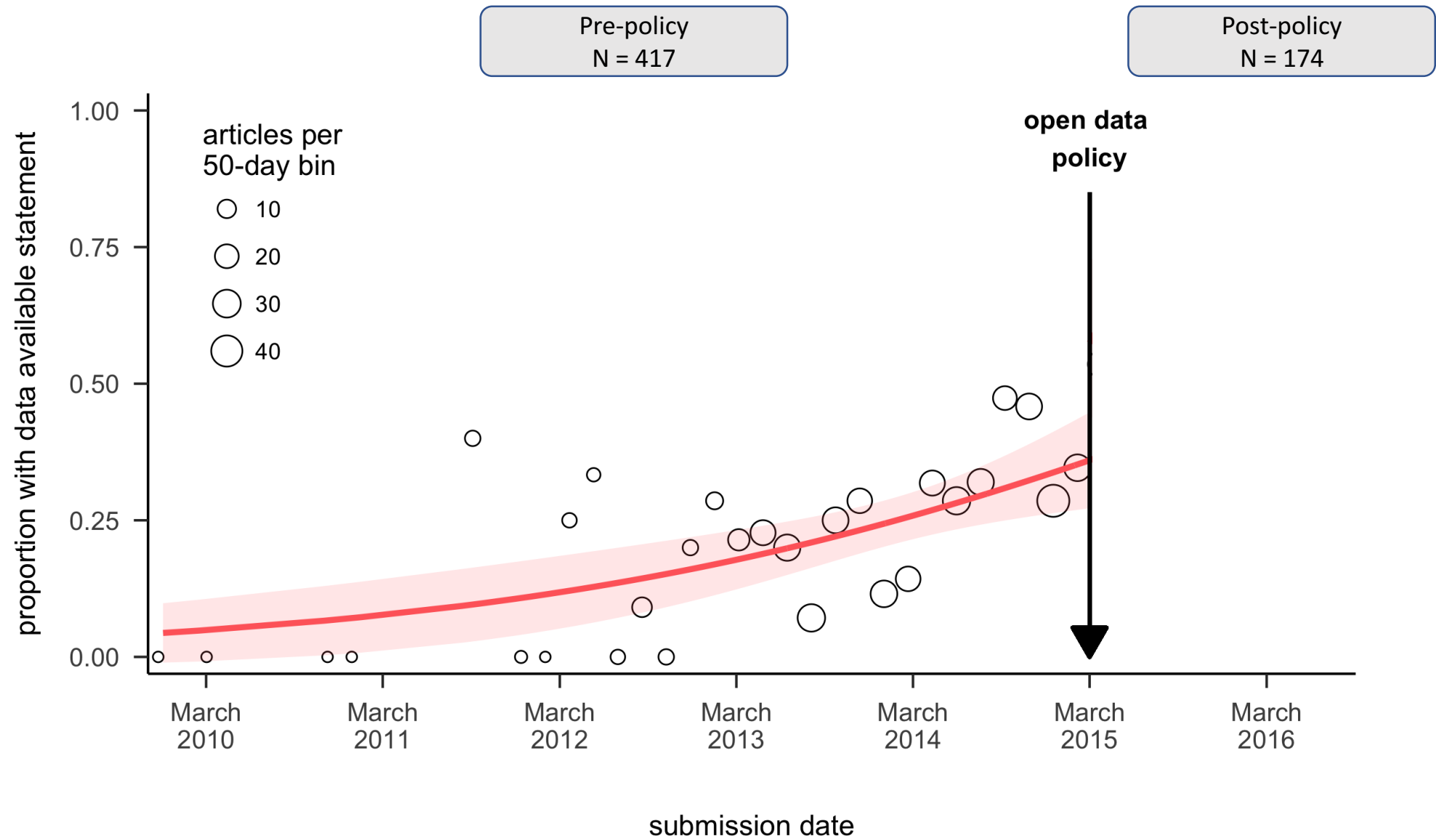
In-principle
reusability
(IPR)





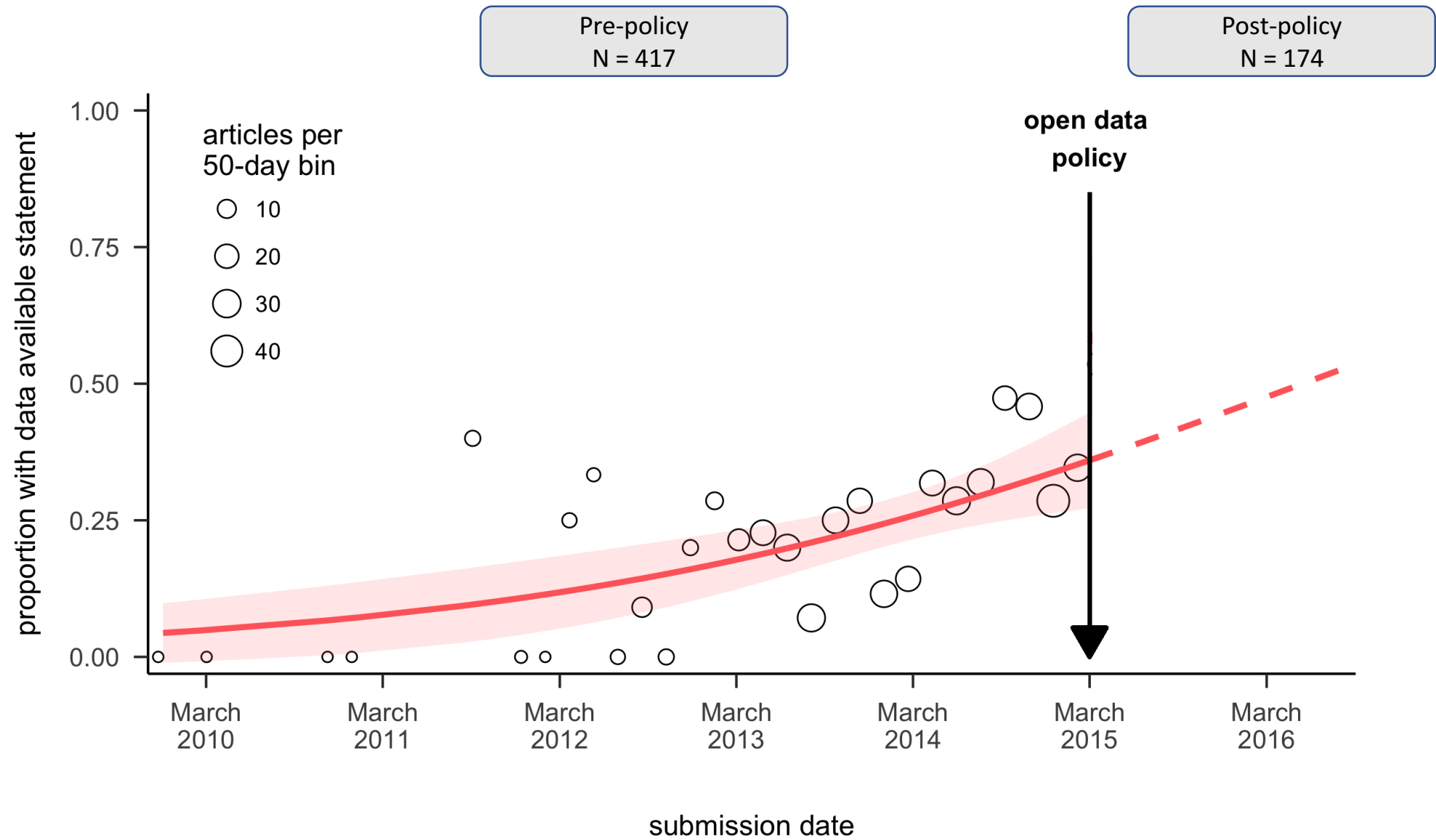
Hardwicke et al. (2018)





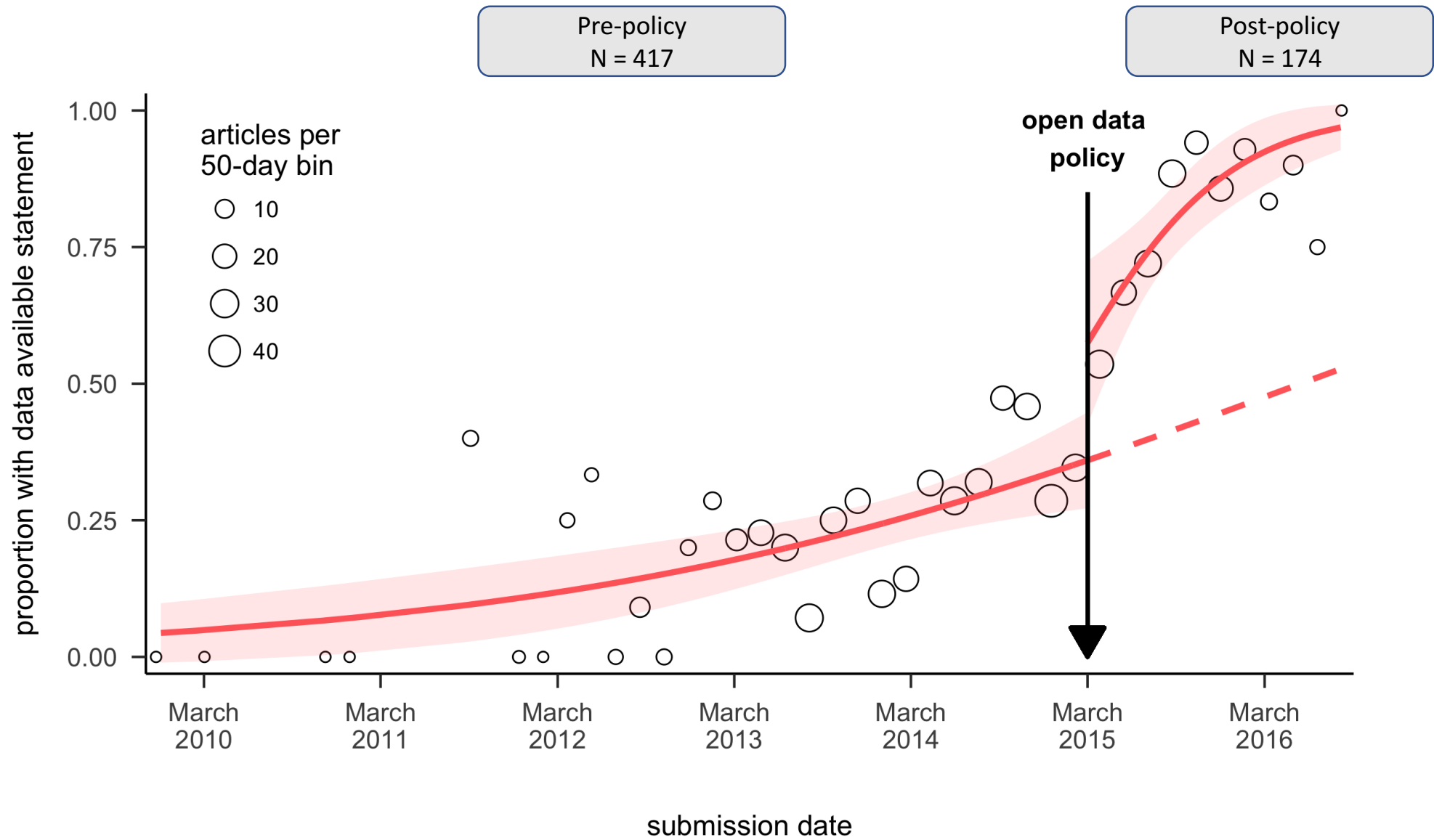
Hardwicke et al. (2018)



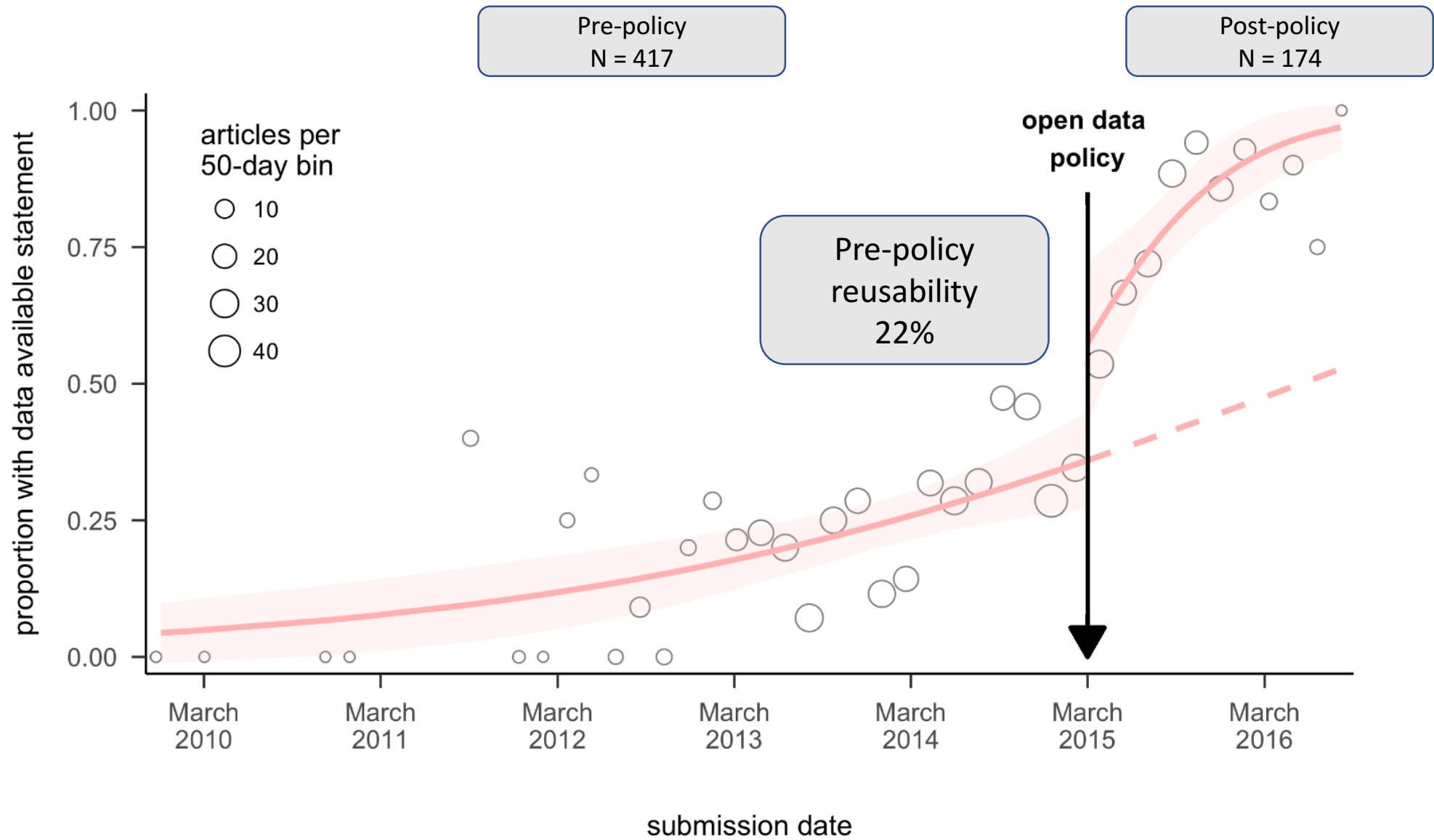


Hardwicke et al. (2018)



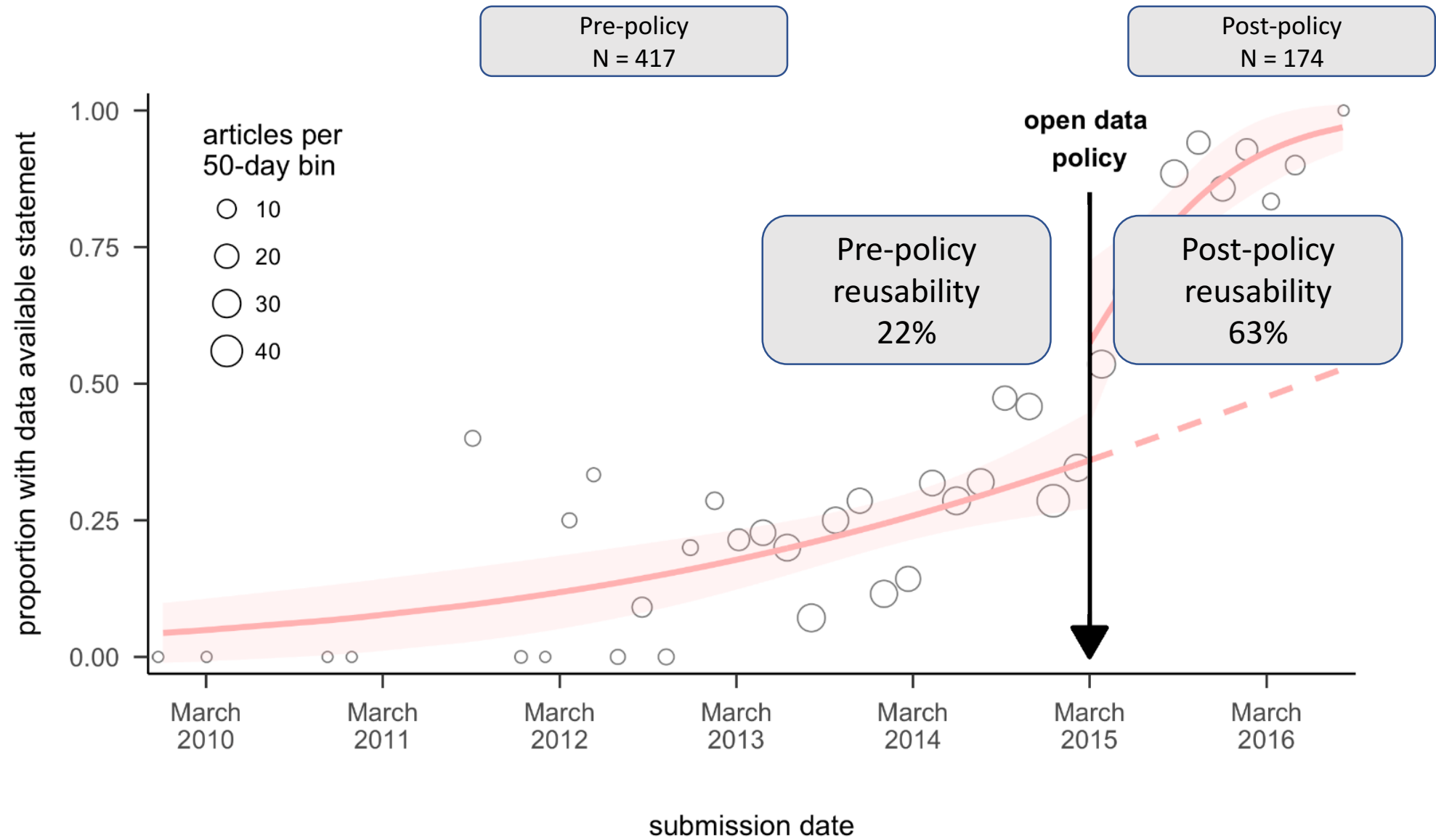


Hardwicke et al. (2018)



Hardwicke et al. (2018)





Hardwicke et al. (2018)





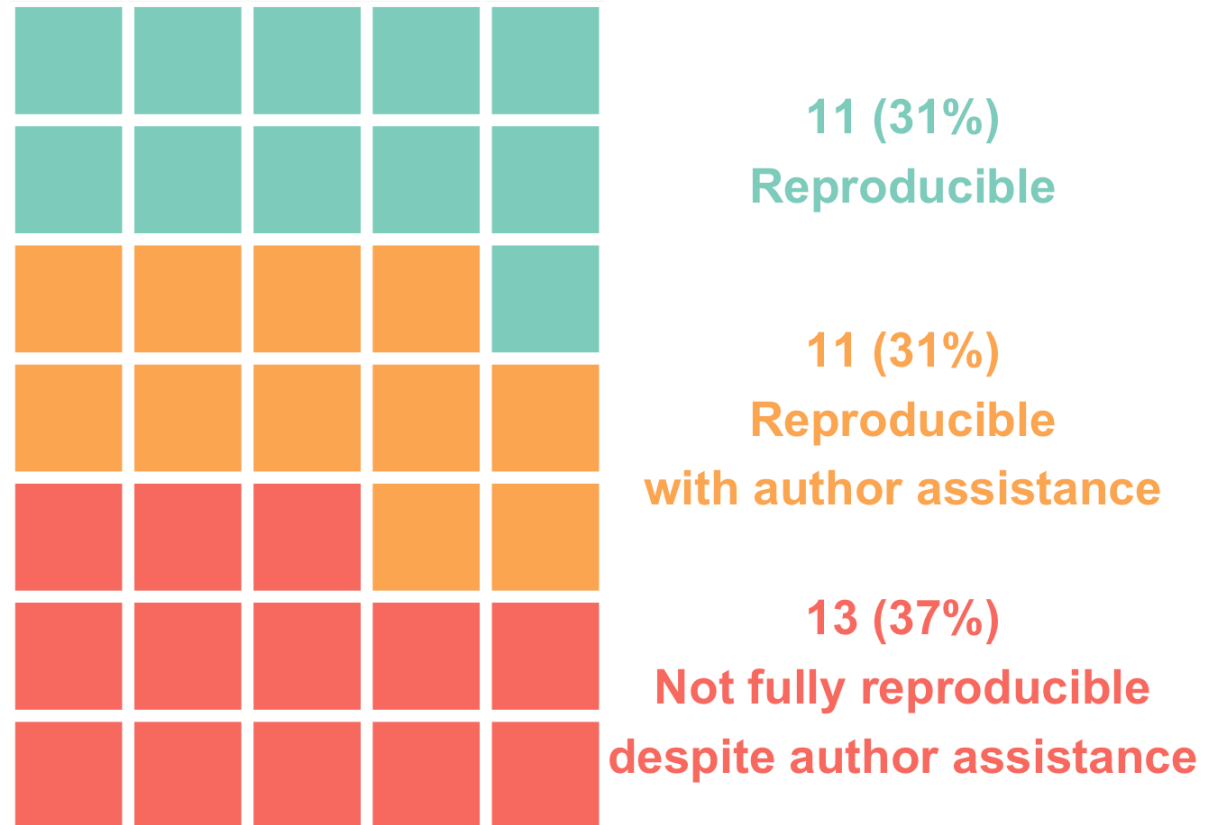
Assessment of analytic reproducibility

- Identified a ‘straightforward’ and substantive finding from a pseudo-randomly selected subset of 35 articles from Stage 1 judged to have re-usable data in principle.
- Analysis “co-piloting” model.
- Always sought author assistance/clarification when issues arose.
- “Major error” is a $\geq 10\%$ discrepancy between reported and obtained value. An article is ‘not fully reproducible’ if it has one or more major errors.

Hardwicke et al. (2018)



Assessment of analytic reproducibility



Hardwicke et al. (2018)

Assessment of analytic reproducibility

Important caveat:

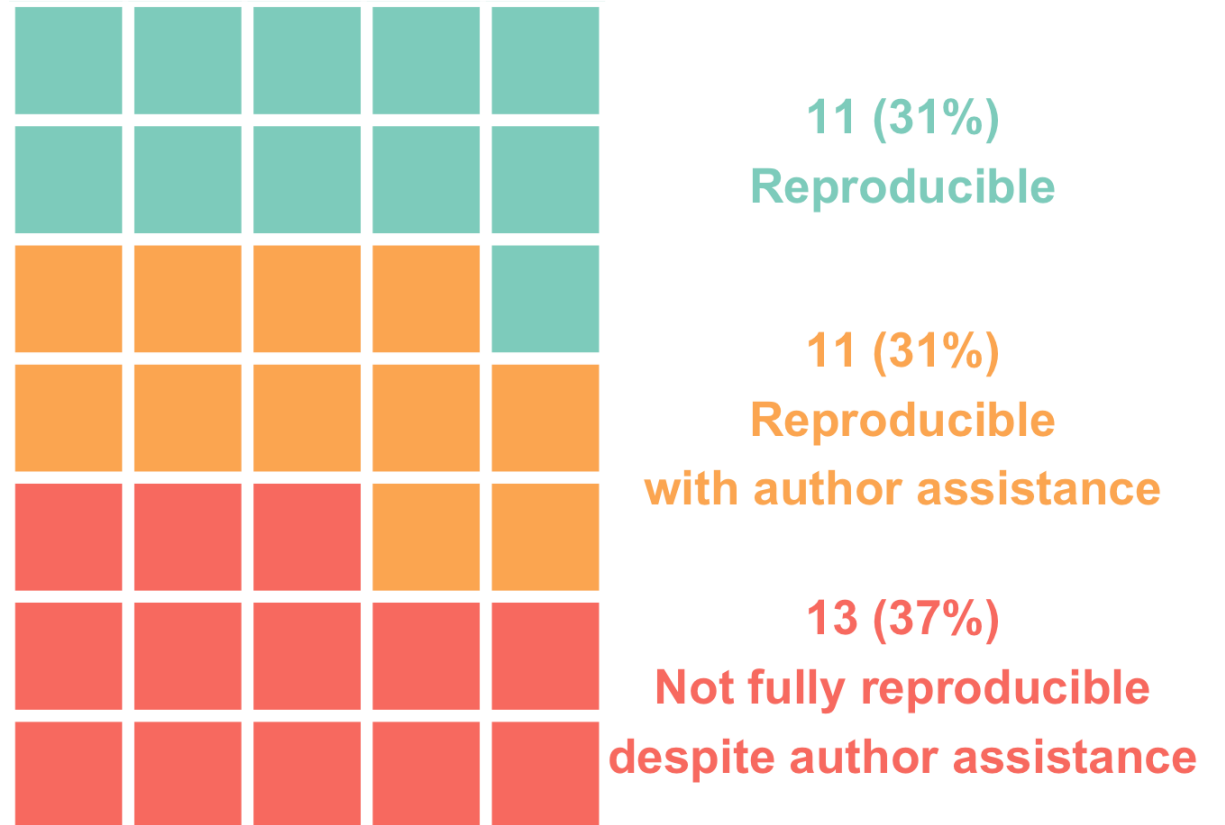
No cases where reproducibility issues appeared to seriously undermine substantive conclusions (3 unclear cases)

Overall error rate:

1324 values checked. 64 “major numerical errors” (5% error rate)

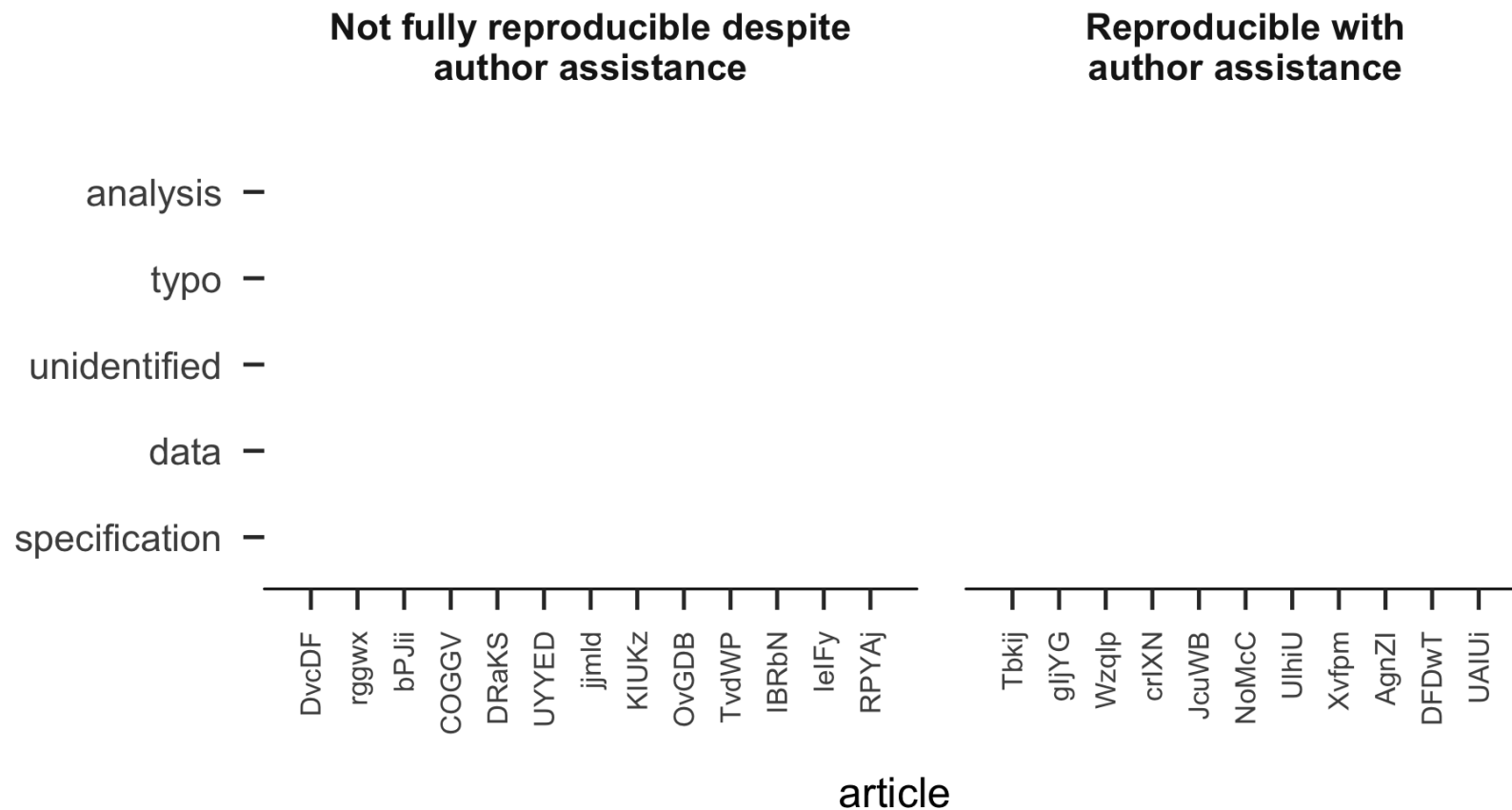
Person hours for each:

2 - 4 (no assistance) or 5 - 25 (assistance)



Hardwicke et al. (2018)

locus of non-reproducibility



issue status:

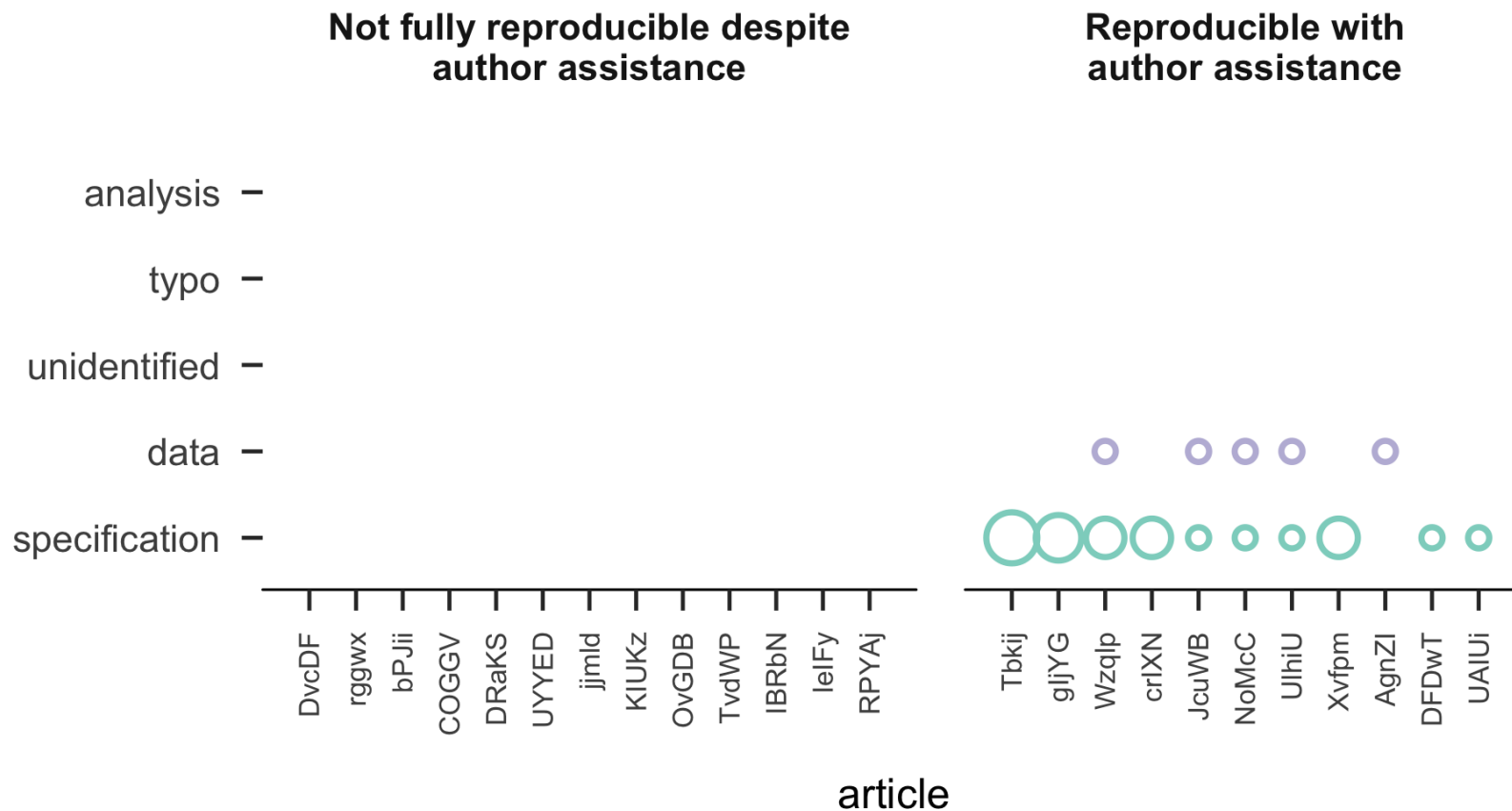
○ resolved ✕ unresolved

issue count:

✕ 1 ✕ 2 ✕ 3 ✕ 4

Hardwicke et al. (2018)

locus of non-reproducibility



issue status:

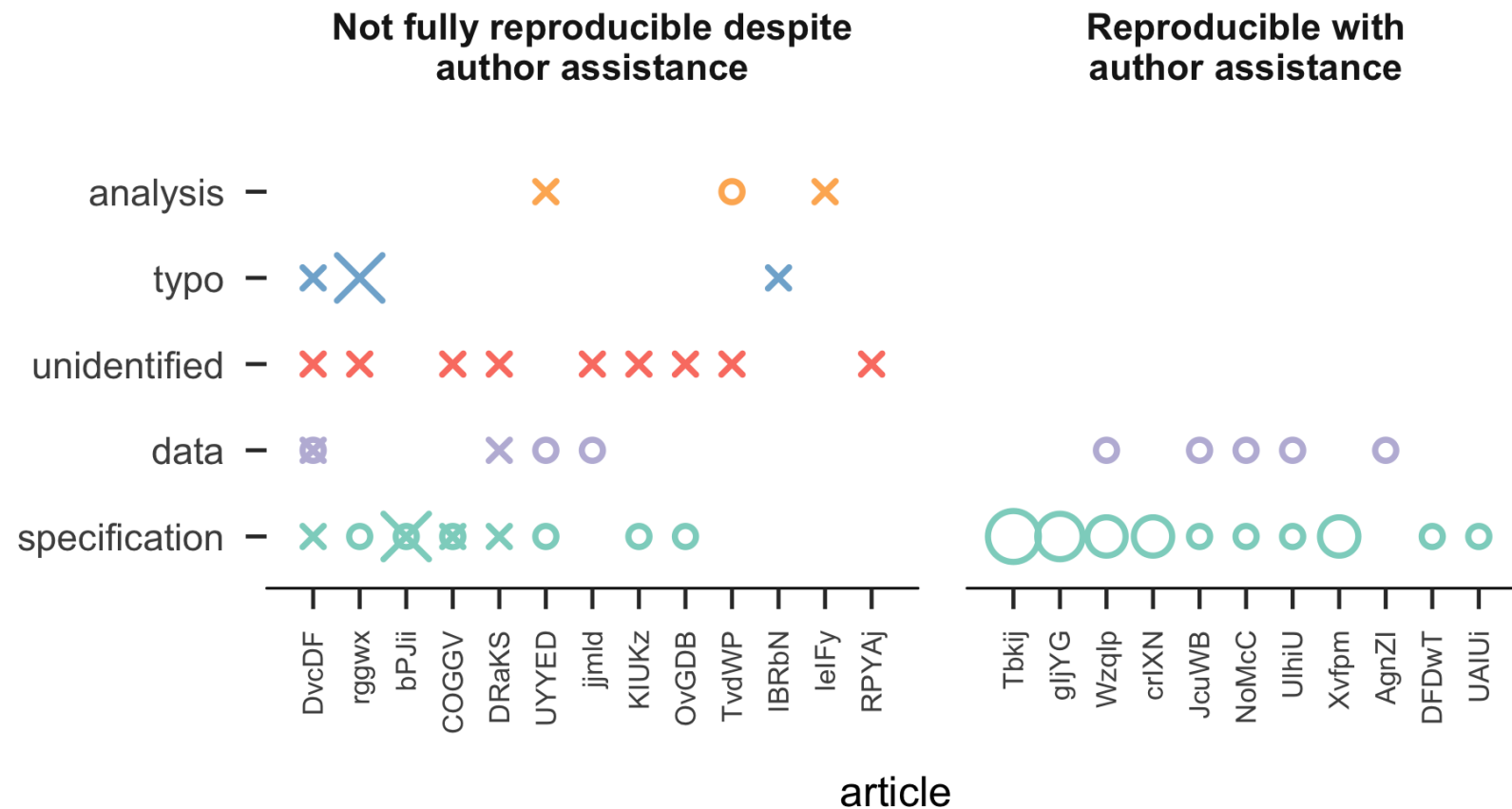
○ resolved ✕ unresolved

issue count:

✕ 1 ✕ 2 ✕ 3 ✕ 4

Hardwicke et al. (2018)

locus of non-reproducibility



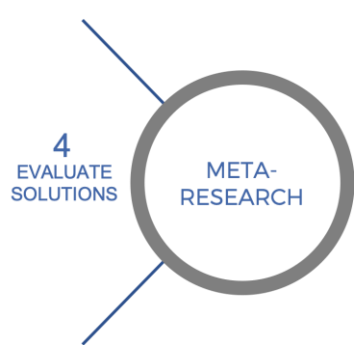
issue status:

○ resolved × unresolved

issue count:

× 1 × 2 × 3 × 4

Hardwicke et al. (2018)



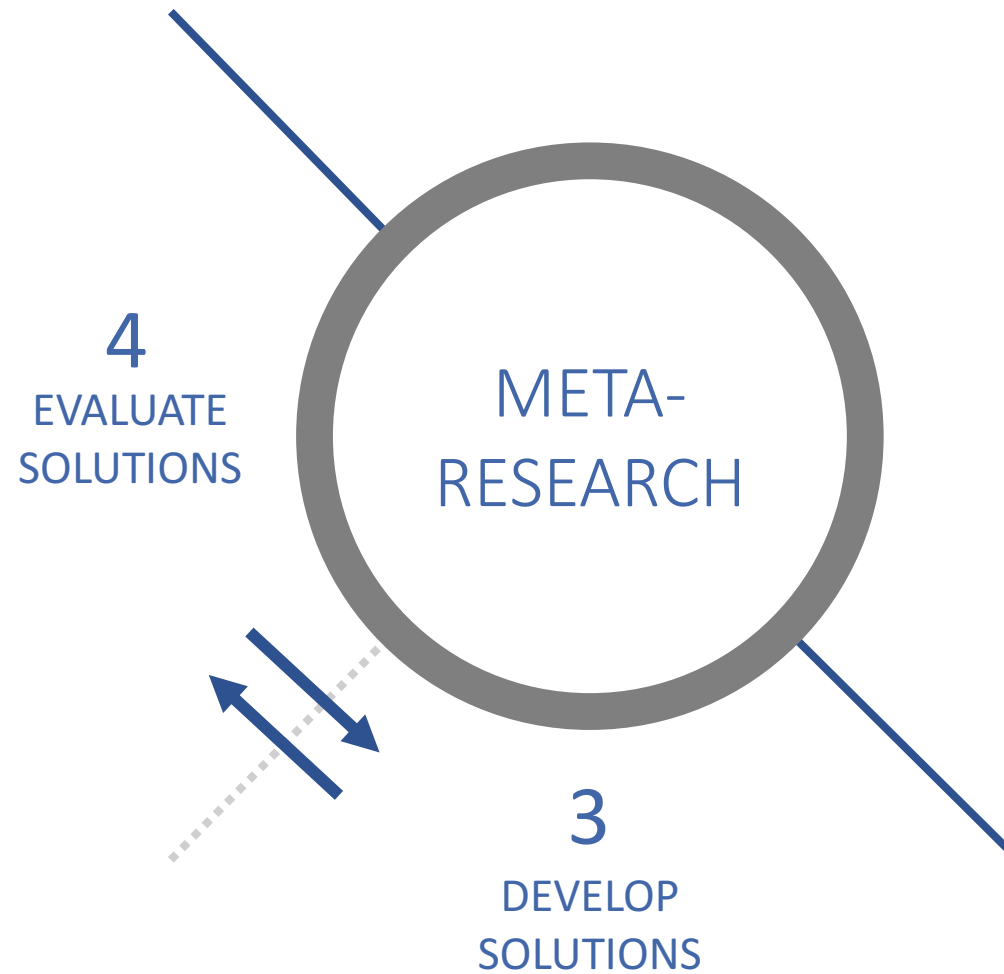
Conclusions

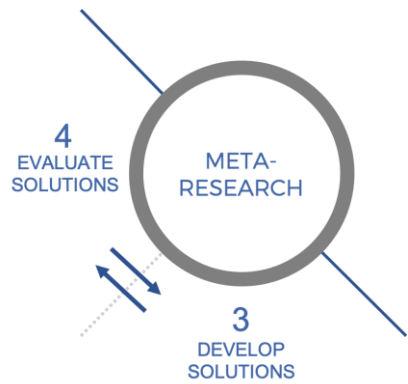
- >> Cognition's mandatory open data policy highly effective, but fell short of ideal.
- >> The initial non-reproducibility of approximately two-thirds of the assessed cases, and the substantial time and effort expended attempting to establish analytic reproducibility, implies a rather serious deterrent to any scientist considering reusing shared data.
- >> It seems likely that our sample is biased towards optimistic reproducibility outcomes: we only evaluated a subset of outcomes based on straightforward analyses and data that was already IPR.
- >> Reproducibility issues may be major barriers to data re-use but do not necessarily undermine substantive conclusions.
- >> The way we write scientific papers is not verifiable and highly error prone – we should move towards writing fully reproducible papers.

Hardwicke et al. (2018)









Meta-research feedback cycles



Contents lists available at [ScienceDirect](#)

Cognition

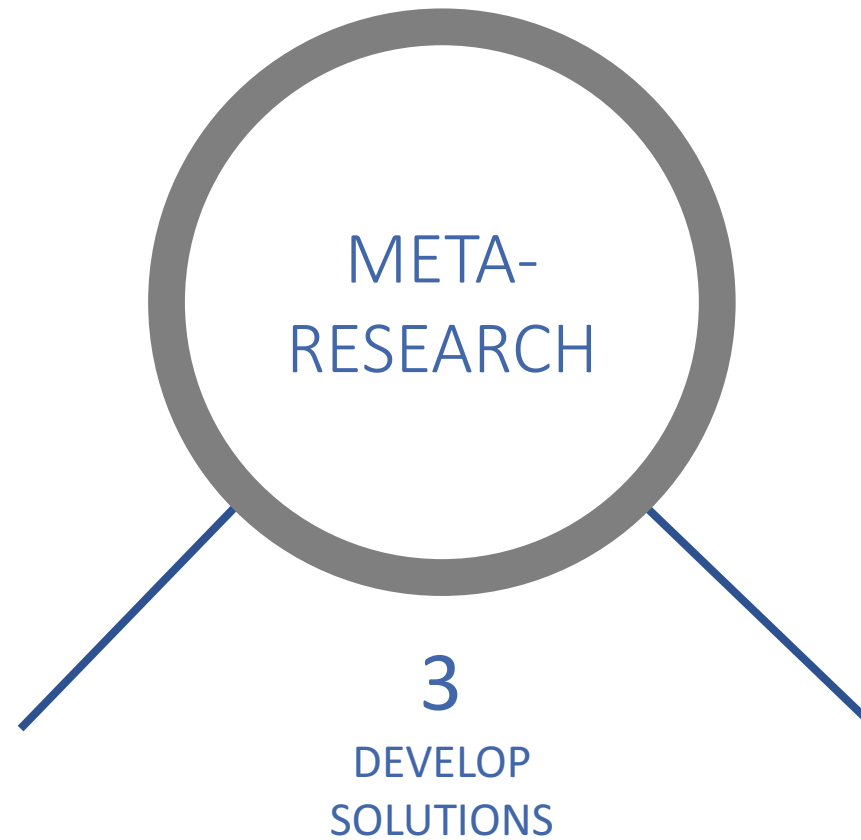
journal homepage: www.elsevier.com/locate/cognit



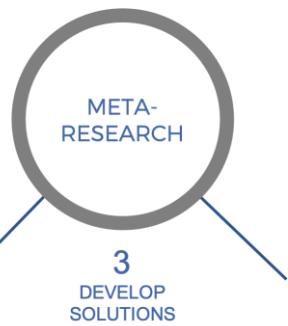
Re-thinking Cognition's Open Data Policy: Responding to Hardwicke and colleagues' evaluation of its impact



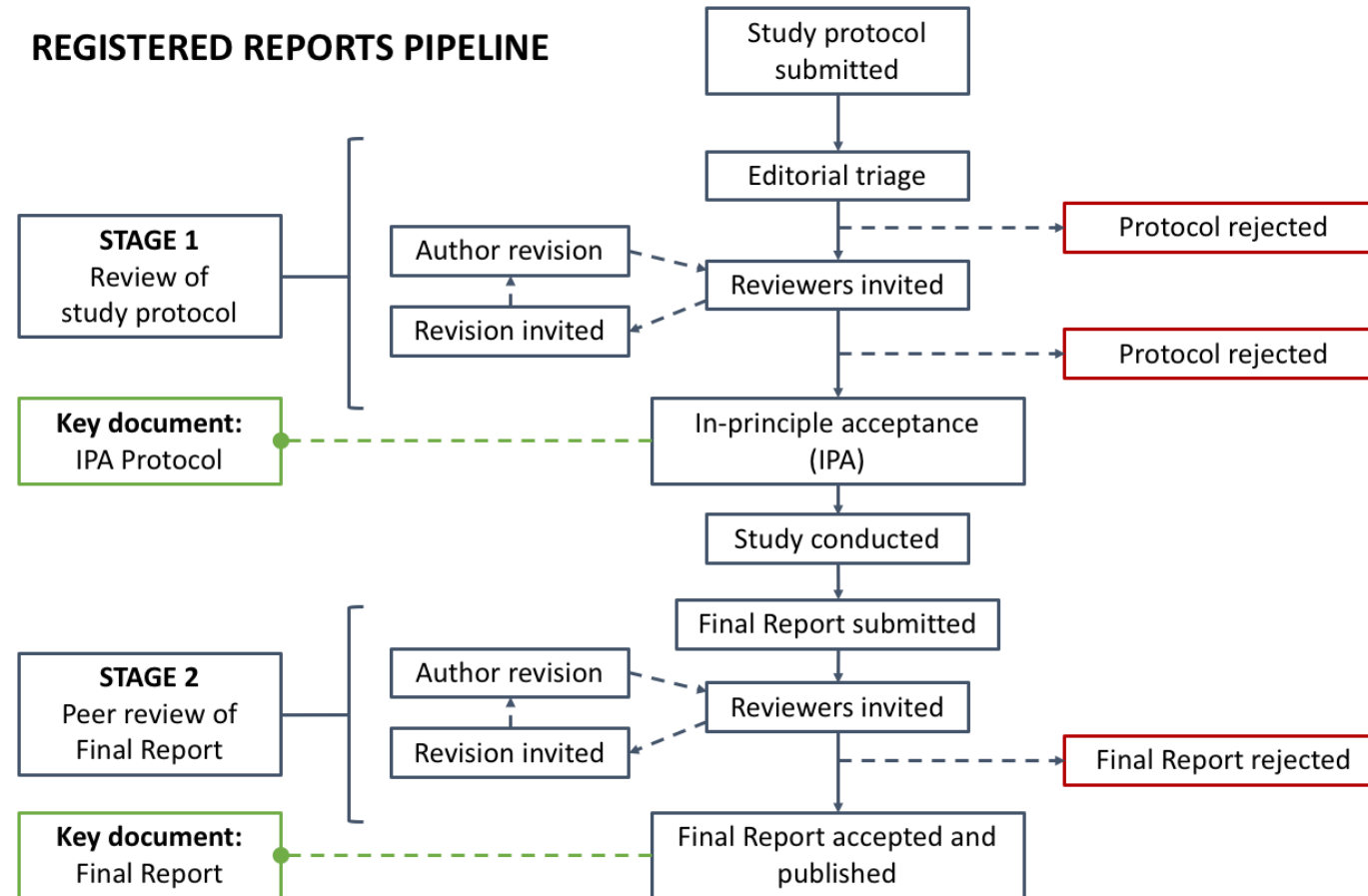


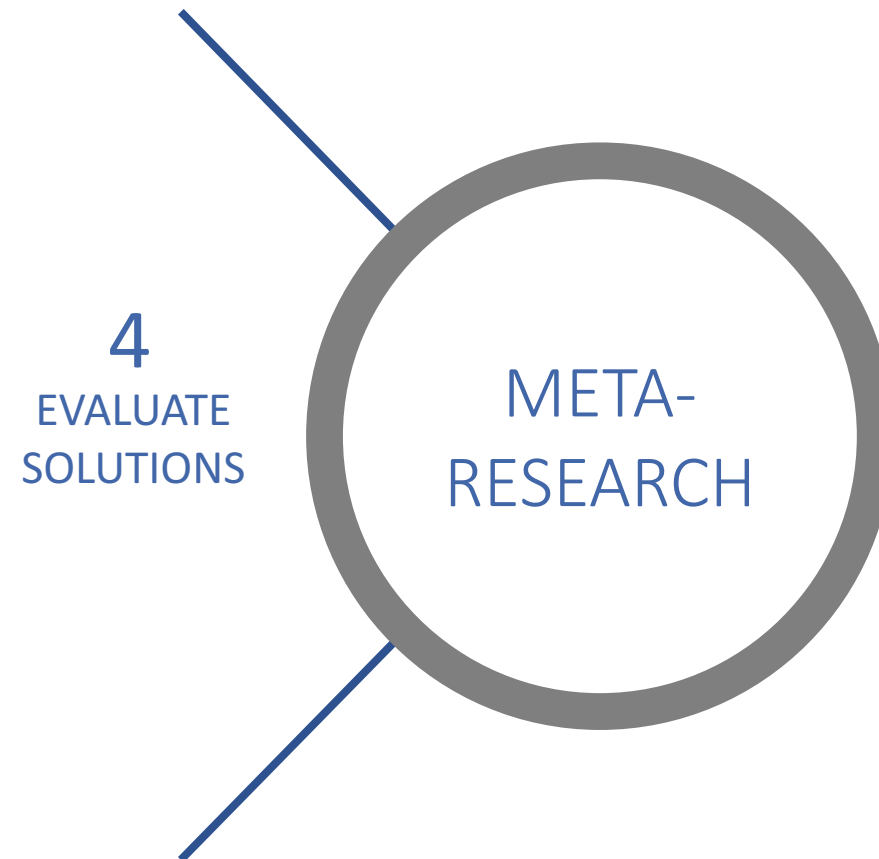


Pre-registration and Registered Reports

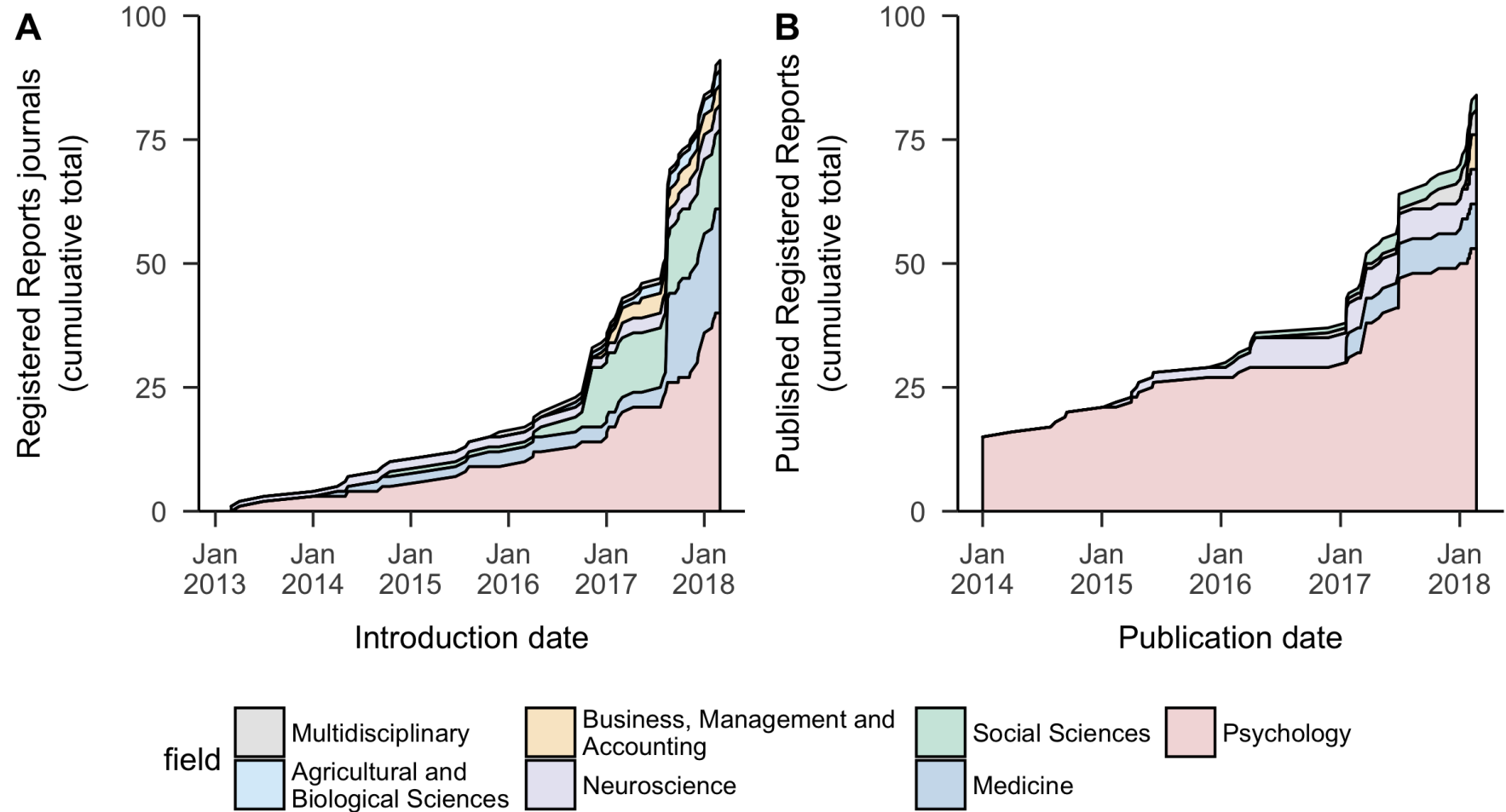


REGISTERED REPORTS PIPELINE





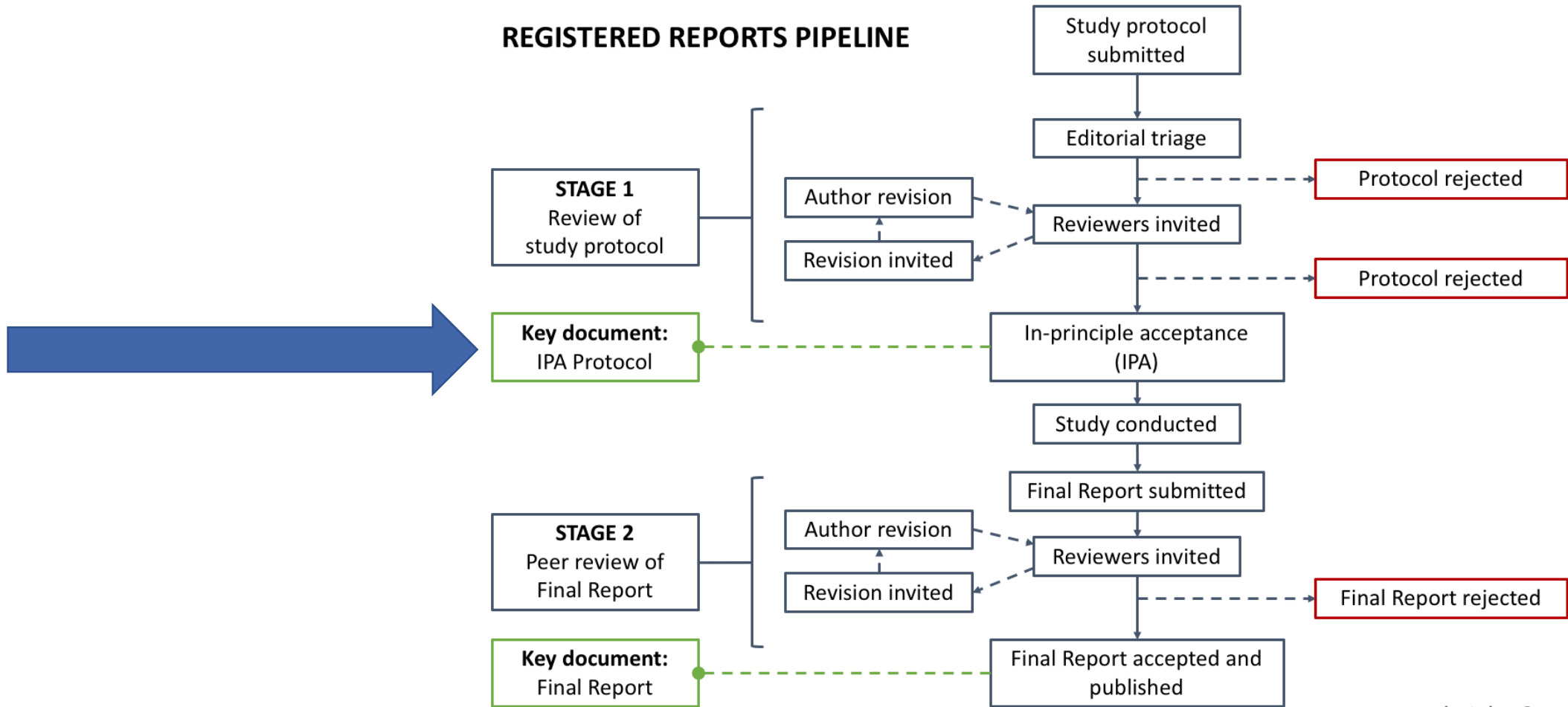
The rise of Registered Reports



Hardwicke & Ioannidis (2018)

Protocol availability

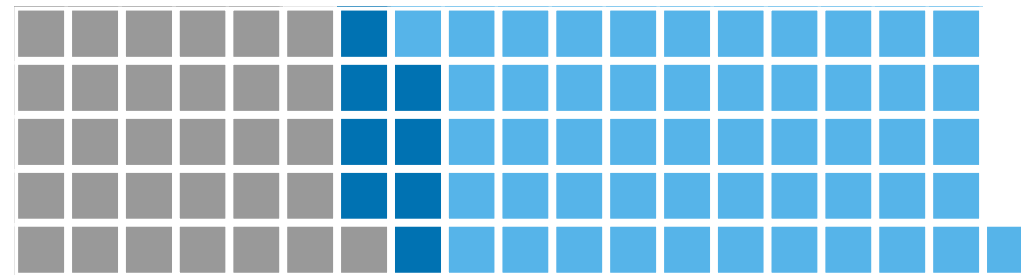
REGISTERED REPORTS PIPELINE



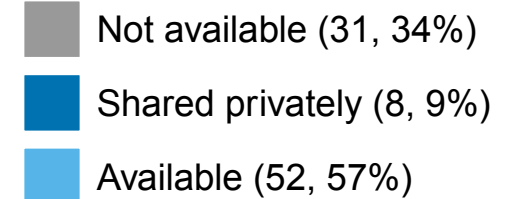
Hardwicke & Ioannidis (2018)

Protocol availability

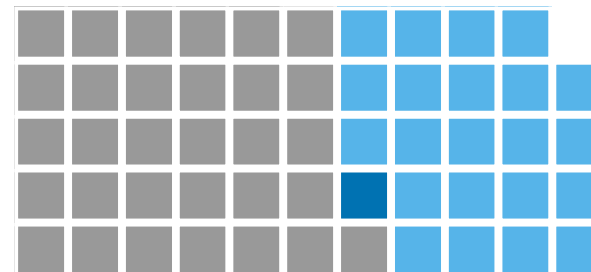
A. Final Report published



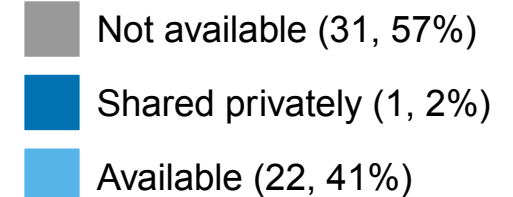
IPA protocol status:



B. Final Report not yet published



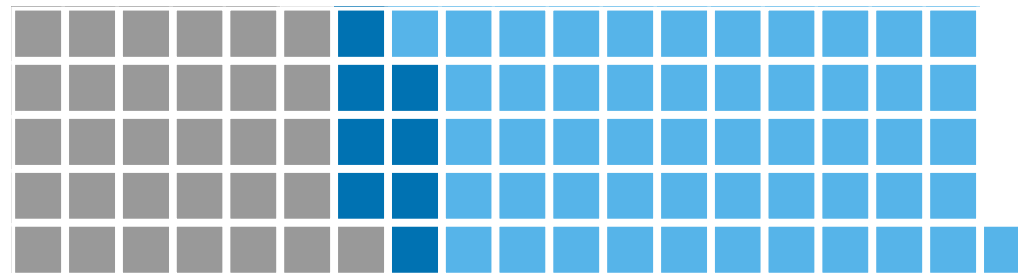
IPA protocol status:



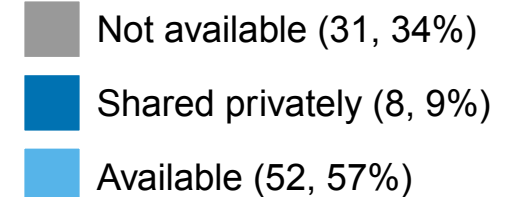
Protocol availability

Of the 74 publicly available IPA protocols, only 26 had been formally registered.

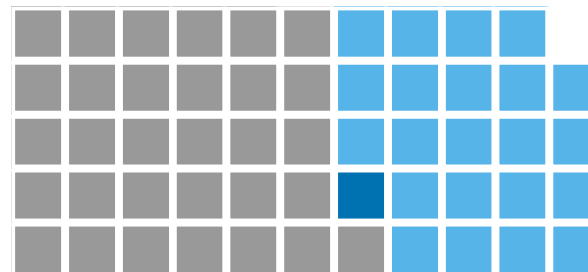
A. Final Report published



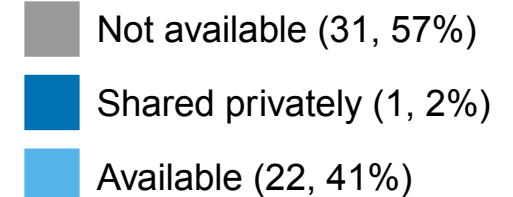
IPA protocol status:



B. Final Report not yet published



IPA protocol status:

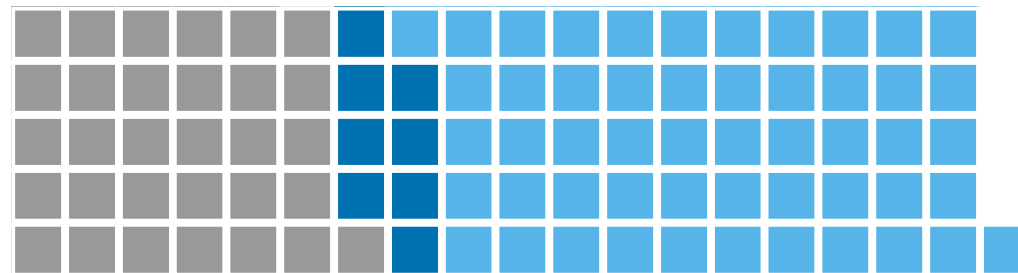


Protocol availability

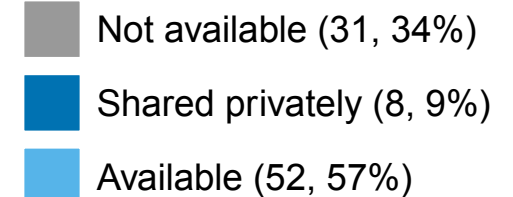
Of the 74 publicly available IPA protocols, only 26 had been formally registered.

Of the 91 published RRs, only 41 self-identified as RRs.

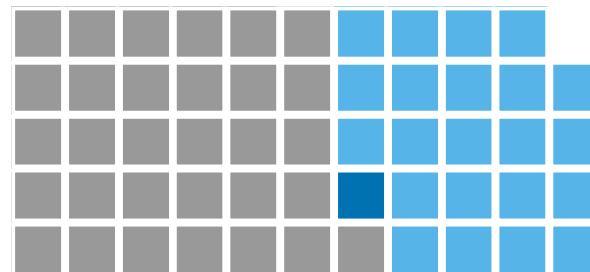
A. Final Report published



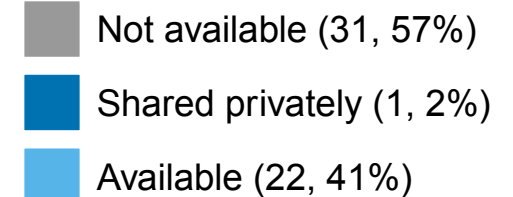
IPA protocol status:



B. Final Report not yet published



IPA protocol status:



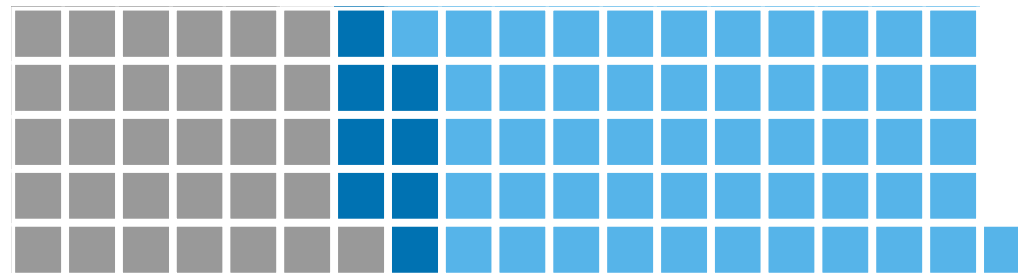
Protocol availability

Of the 74 publicly available IPA protocols, only 26 had been formally registered.

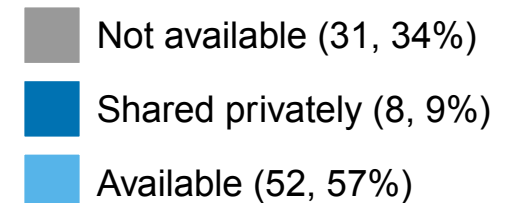
Of the 91 published RRs, only 41 self-identified as RRs.

No reliable way of tracking status in publication pipeline.

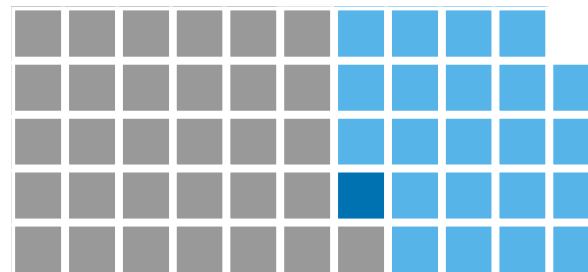
A. Final Report published



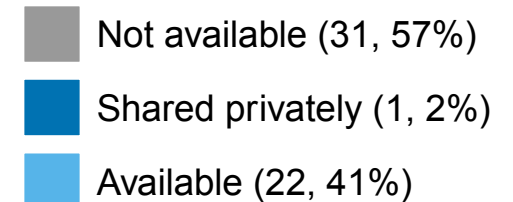
IPA protocol status:

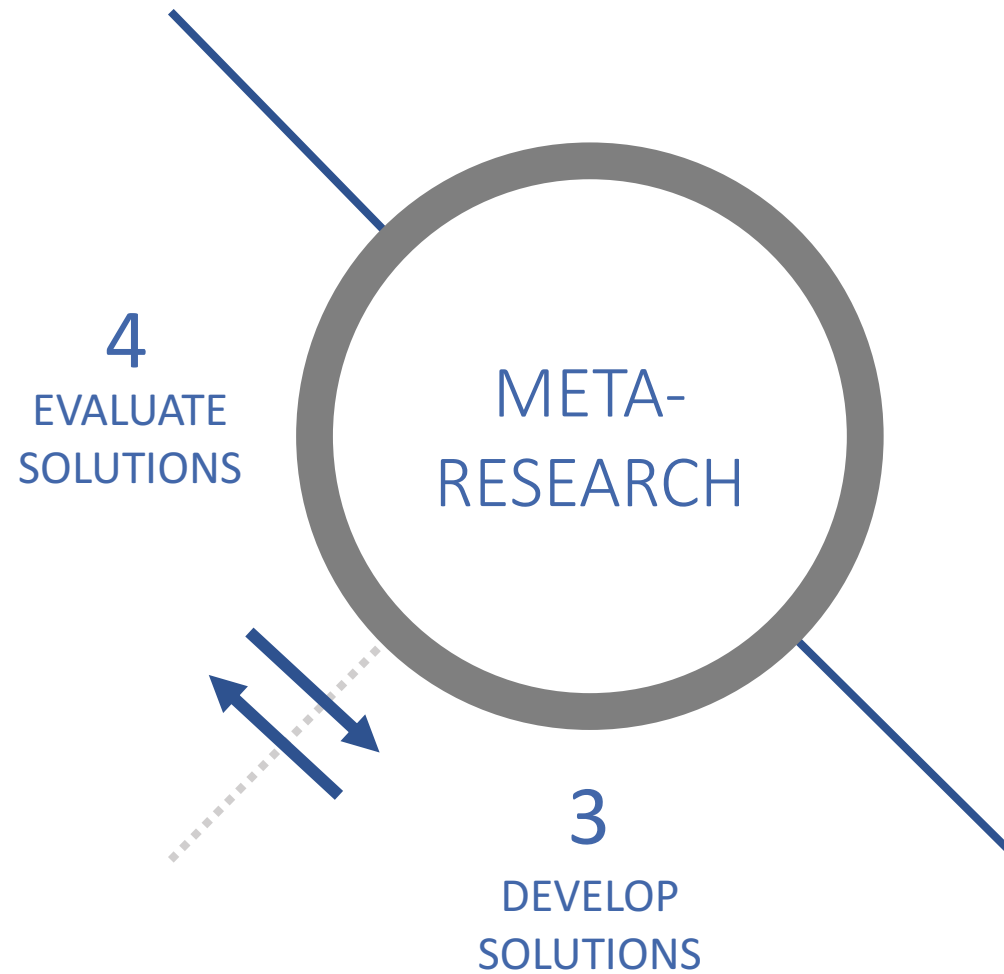


B. Final Report not yet published

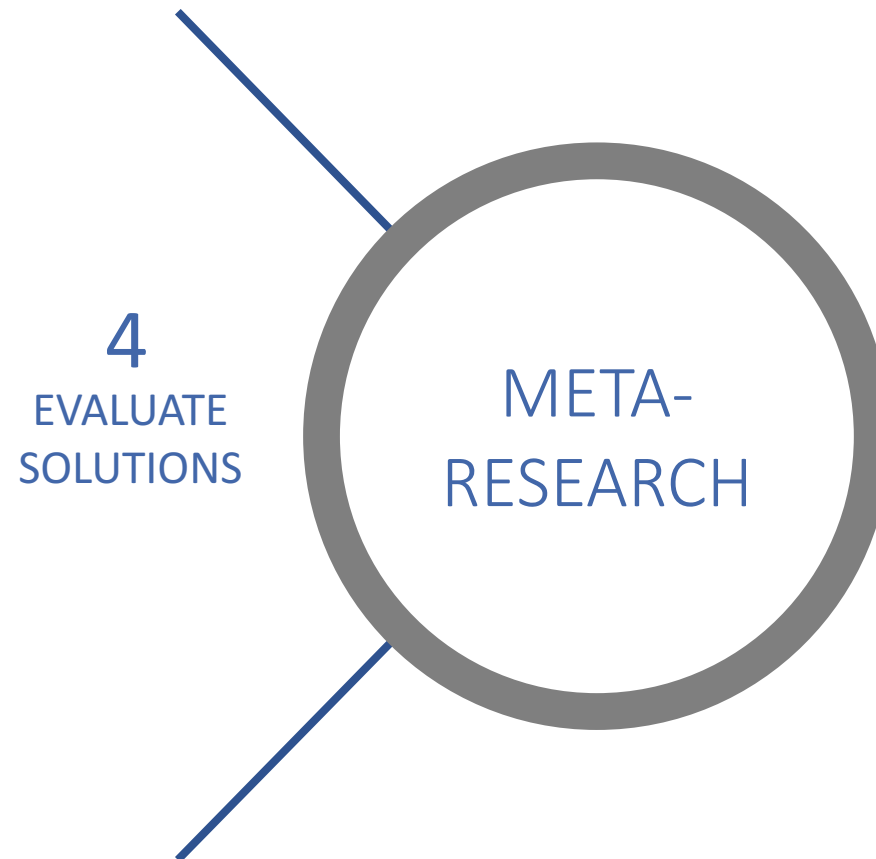


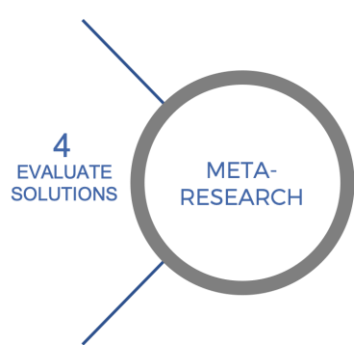
IPA protocol status:











Pre-registration problems



COMPARE

TRACKING SWITCHED OUTCOMES IN CLINICAL TRIALS

67

TRIALS CHECKED

9

TRIALS WERE
PERFECT

354

OUTCOMES NOT
REPORTED

357

NEW OUTCOMES
SILENTLY ADDED

Use of Trial Register Information during the Peer Review Process

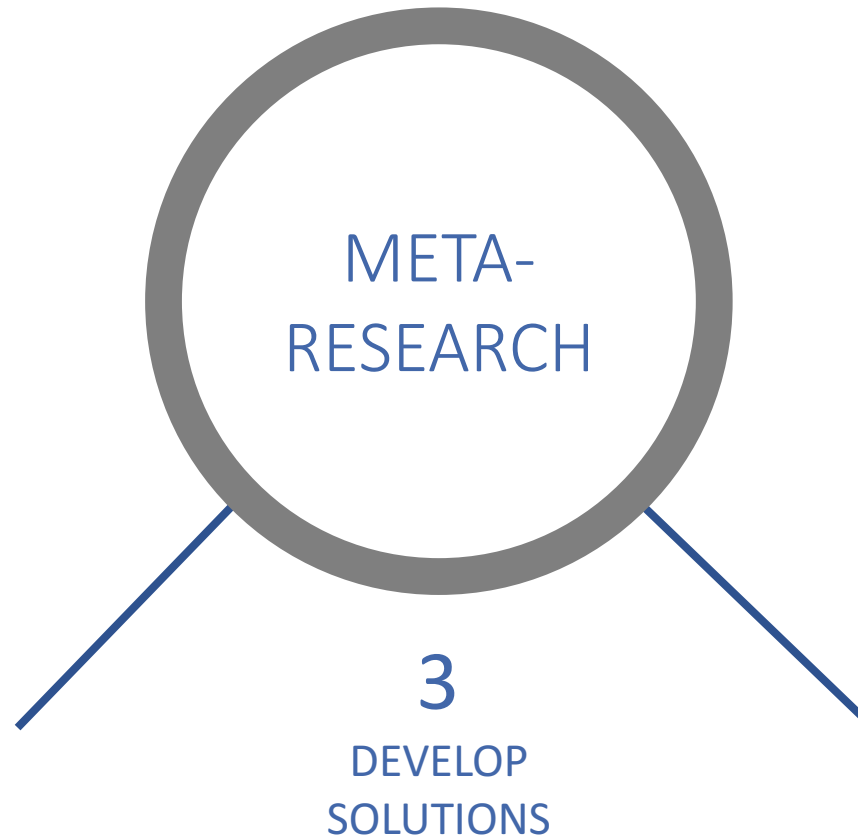
Sylvain Mathieu^{1,2}, An-Wen Chan³, Philippe Ravaud^{1*}

¹INSERM U738, Centre d'épidémiologie Clinique, French Cochrane Center, University Paris Descartes et Hotel Dieu, Paris, France, ²Department of Rheumatology, University Clermont 1, Clermont-Ferrand, France, ³Women's College Research Institute, University of Toronto, Toronto, Ontario, Canada

232/672 (34.3%) of clinical trial reviewers said that they had examined the registered protocol.



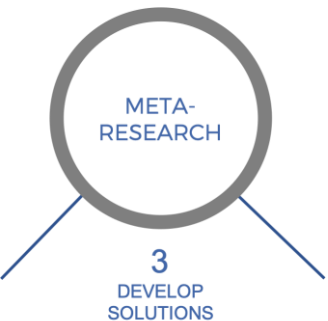
META-RESEARCH INNOVATION
CENTER BERLIN



SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously





SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously

STEP 1: Add structured markup to pre-registered protocols.

EXAMPLE:

Pre-registered
decision units
(PDUs)

P-1: We hypothesise that participants in the drug condition will have better mood scores than participants in the placebo control condition.

P-2: The dependent variable is mood scores as measured on a 5-point Likert scale.

....

P-6: A power analysis was carried out to determine that a sample size of 61 participants is necessary to detect an effect of $f = .25$ with power of .9.



SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously

STEP 1: Add structured markup to pre-registered protocols.

EXAMPLE:

Unique tag
with XML
anchor

Pre-registered
decision units
(PDUs)

P-1: We hypothesise that participants in the drug condition will have better mood scores than participants in the placebo control condition.

P-2: The dependent variable is mood scores as measured on a 5-point Likert scale.

....

P-6: A power analysis was carried out to determine that a sample size of 61 participants is necessary to detect an effect of $f = .25$ with power of .9.



SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously

META-RESEARCH

3

DEVELOP
SOLUTIONS

STEP 1: Add structured markup to pre-registered protocols.

Unique tag
with XML
anchor

[optional]
semantic
enrichment

EXAMPLE:

Pre-registered
decision units
(PDUs)

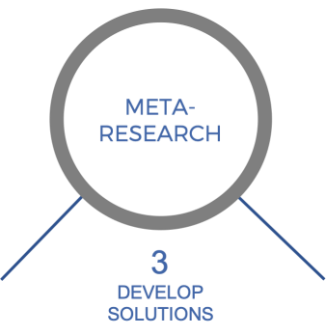
PH-1: We hypothesise that participants in the drug condition will have better mood scores than participants in the placebo control condition.

PM-2: The dependent variable is mood scores as measured on a 5-point Likert scale.

....

PA-6: A power analysis was carried out to determine that a sample size of 61 participants is necessary to detect an effect of $f = .25$ with power of .9.

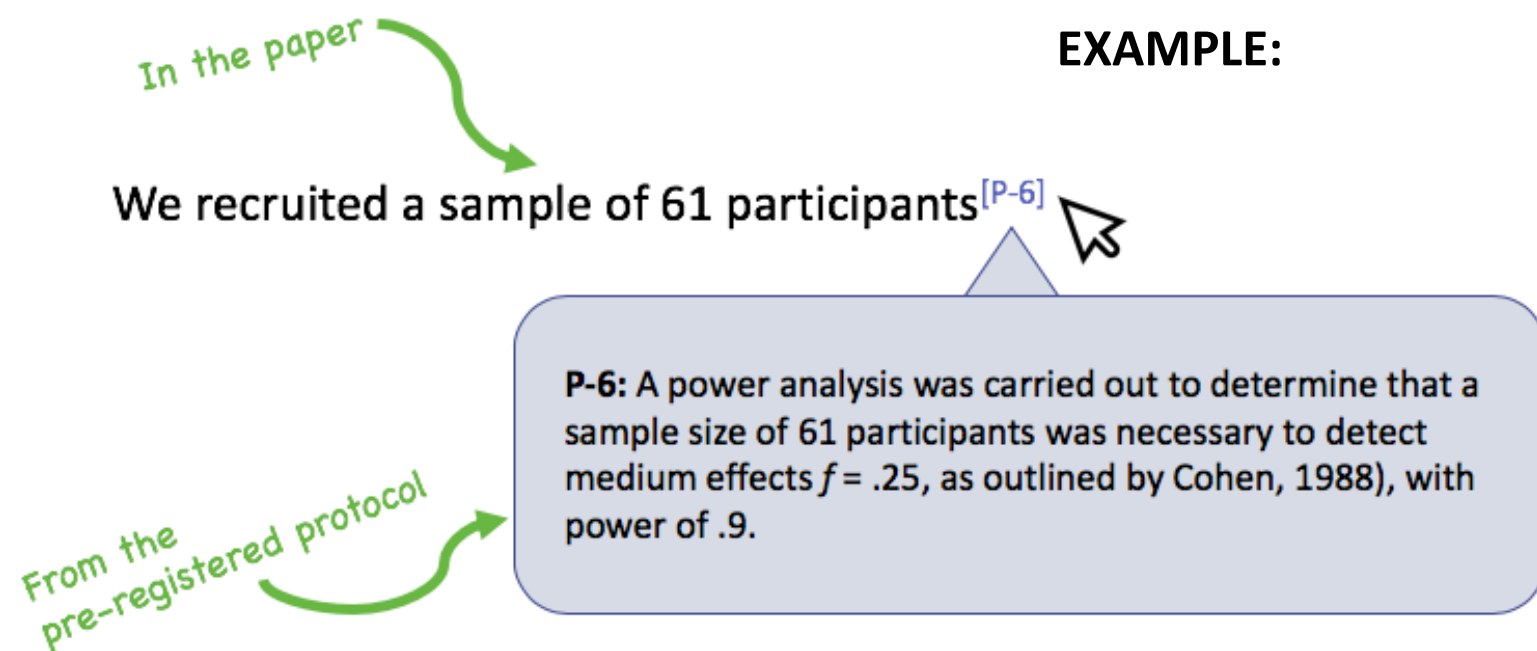


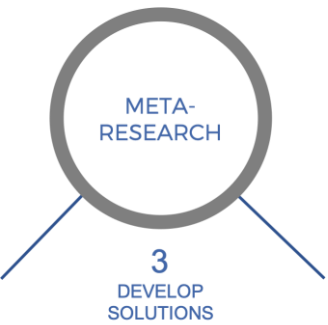


SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously

STEP 2: Add dynamic citation functionality to research papers to facilitate protocol-paper comparison during peer review.





SMART Pre-registration

Sensible Markup Assists Reviewers Tremendously

Potential advantages:

- Granular reporting more accurately conveys pre-registration status.
- Unique PDU tags facilitates verification.
- Side-by-side comparison reduces peer review barriers.
- Automated tools can help to identify selective reporting.



Calibrating the scientific ecosystem through meta-research

1. Scientists are only human. Let's calibrate the scientific ecosystem so that it mitigates our inherent flaws and limitations rather than exacerbates them.
2. Discussion about problems can be enriched with empirical evidence from meta-research.
3. Ongoing monitoring and refinement of proposed solutions can be achieved through meta-research feedback loops.



Acknowledgements



Acknowledgements



John Ioannidis



Steven Goodman



Michael Frank

