The value of openness

Readings for today

- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. Science translational medicine, 8(341), 341ps12-341ps12.
- Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. PLoS Comput Biol, 9(10), e1003285.

Supplemental reading: Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. Annals of the New York Academy of Sciences, 1396(1), 5.

Topics

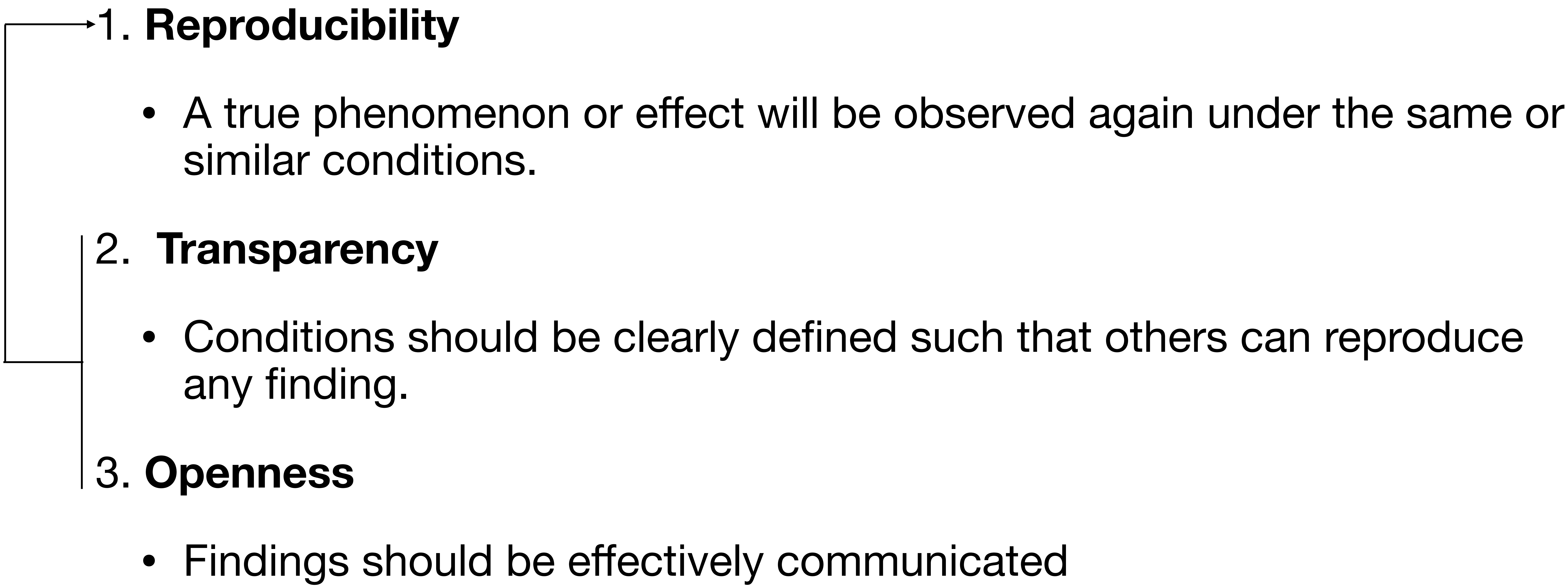
1. Types of reproducibility

2. Open science

3. Rules for reproducible data science

Types of reproducibility

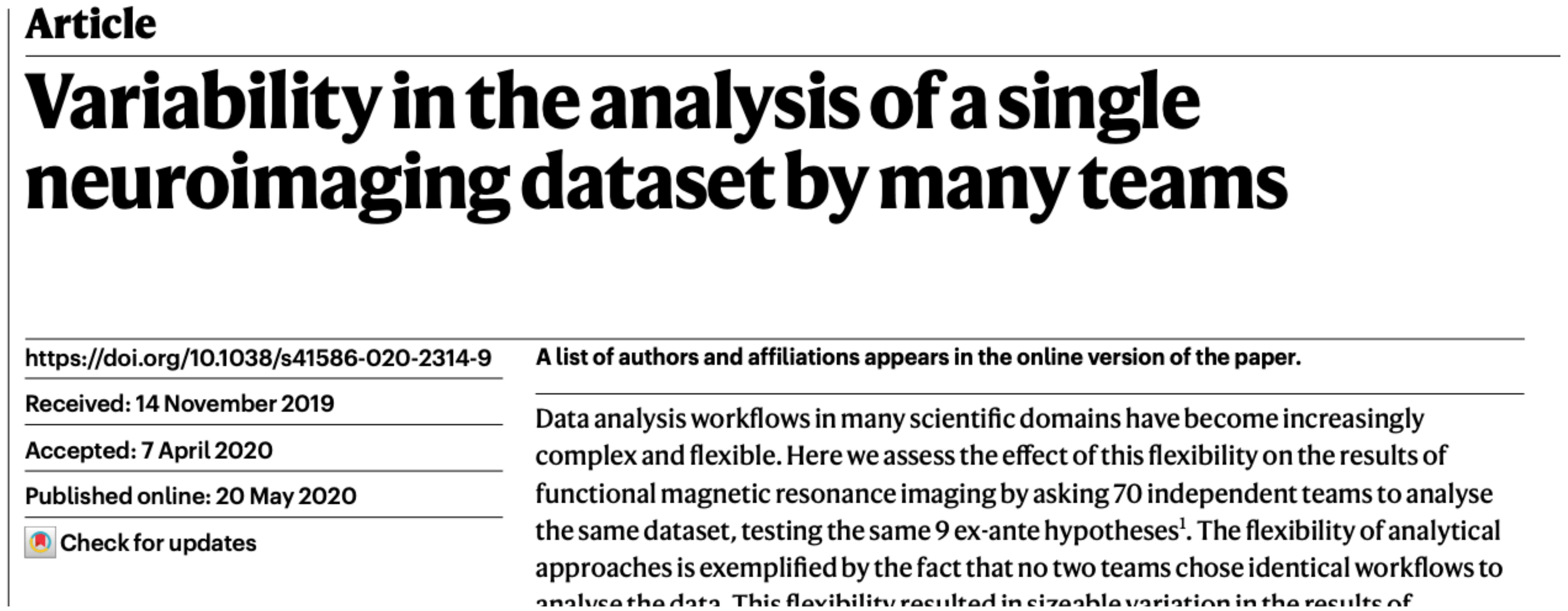
Ideals of science



1. **Reproducibility**
    - A true phenomenon or effect will be observed again under the same or similar conditions.
2. **Transparency**
    - Conditions should be clearly defined such that others can reproduce any finding.
3. **Openness**
    - Findings should be effectively communicated

Types of reproducibility

Problem: Variation in pipelines leads to variation in results.



Article

Variability in the analysis of a single neuroimaging dataset by many teams

https://doi.org/10.1038/s41586-020-2314-9

Received: 14 November 2019

Accepted: 7 April 2020

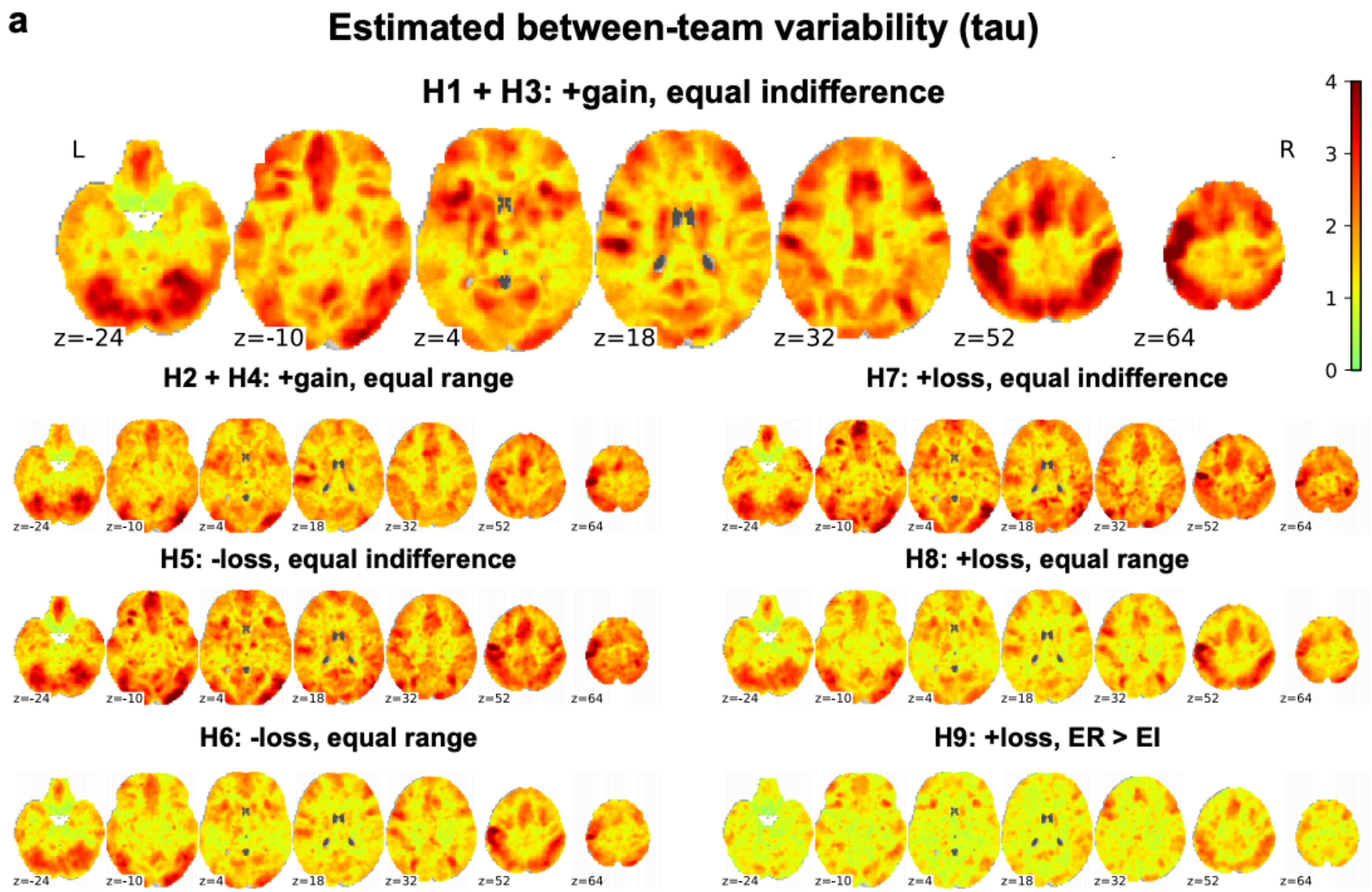
Published online: 20 May 2020

Check for updates

A list of authors and affiliations appears in the online version of the paper.

Data analysis workflows in many scientific domains have become increasingly complex and flexible. Here we assess the effect of this flexibility on the results of functional magnetic resonance imaging by asking 70 independent teams to analyse the same dataset, testing the same 9 ex-ante hypotheses[1]. The flexibility of analytical approaches is exemplified by the fact that no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of

(Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... & Avesani, P. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 1-7.)


a
Estimated between-team variability (tau)
H1 + H3: +gain, equal indifference
L
R
z=-24
z=-10
z=4
z=18
z=32
z=52
z=64
4
3
2
1
0
H2 + H4: +gain, equal range
H7: +loss, equal indifference
H5: -loss, equal indifference
H8: +loss, equal range
H6: -loss, equal range
H9: +loss, ER > EI

(Goodman et al. 2016)

Types of reproducibility

Problem: Variation in pipelines leads to variation in results.

Methods Reproducibility:

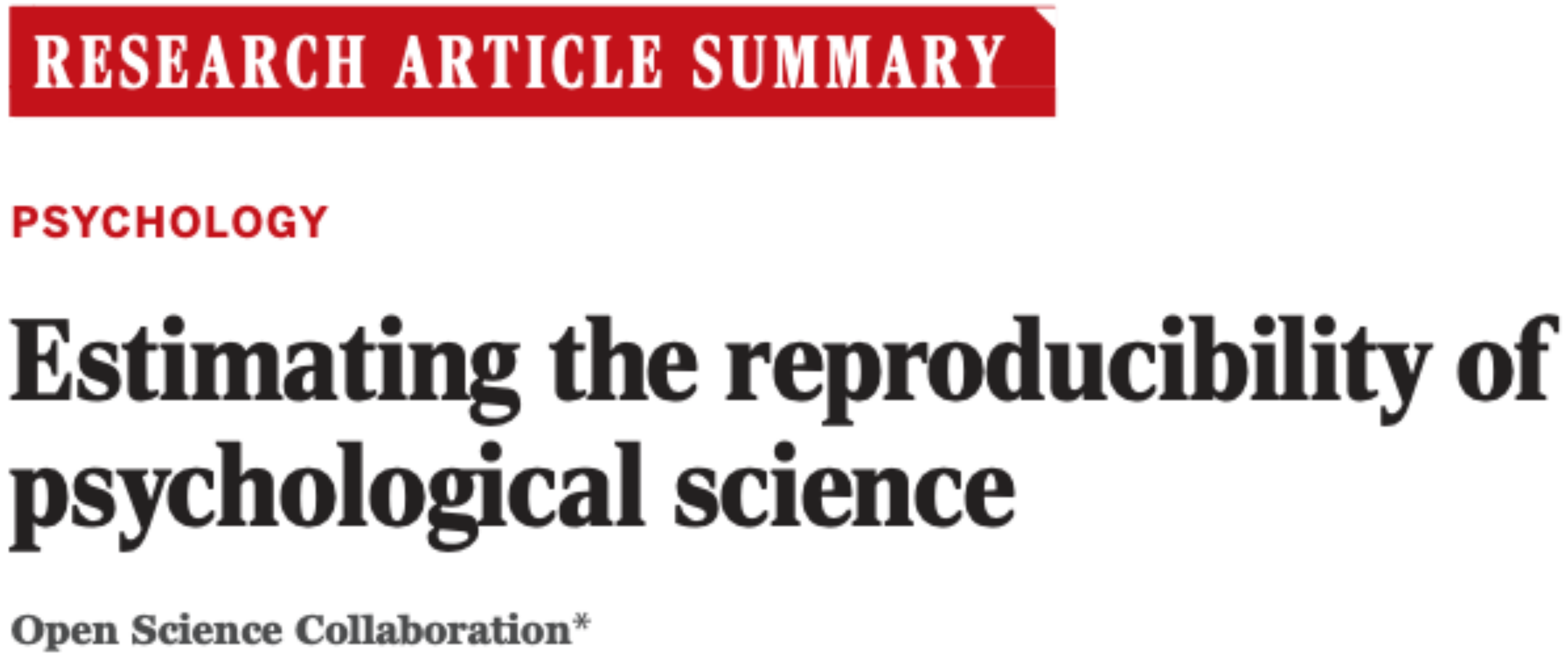
Independent investigators obtain the same results when using the same methods (e.g., tools, analyses) as used in a previous study.

Solution: Share every step of your process from raw data to final figures.

(Goodman et al. 2016)

Types of reproducibility

Problem: Uncontrolled sources of variability can lead to dramatically different finding, even when using identical methods.



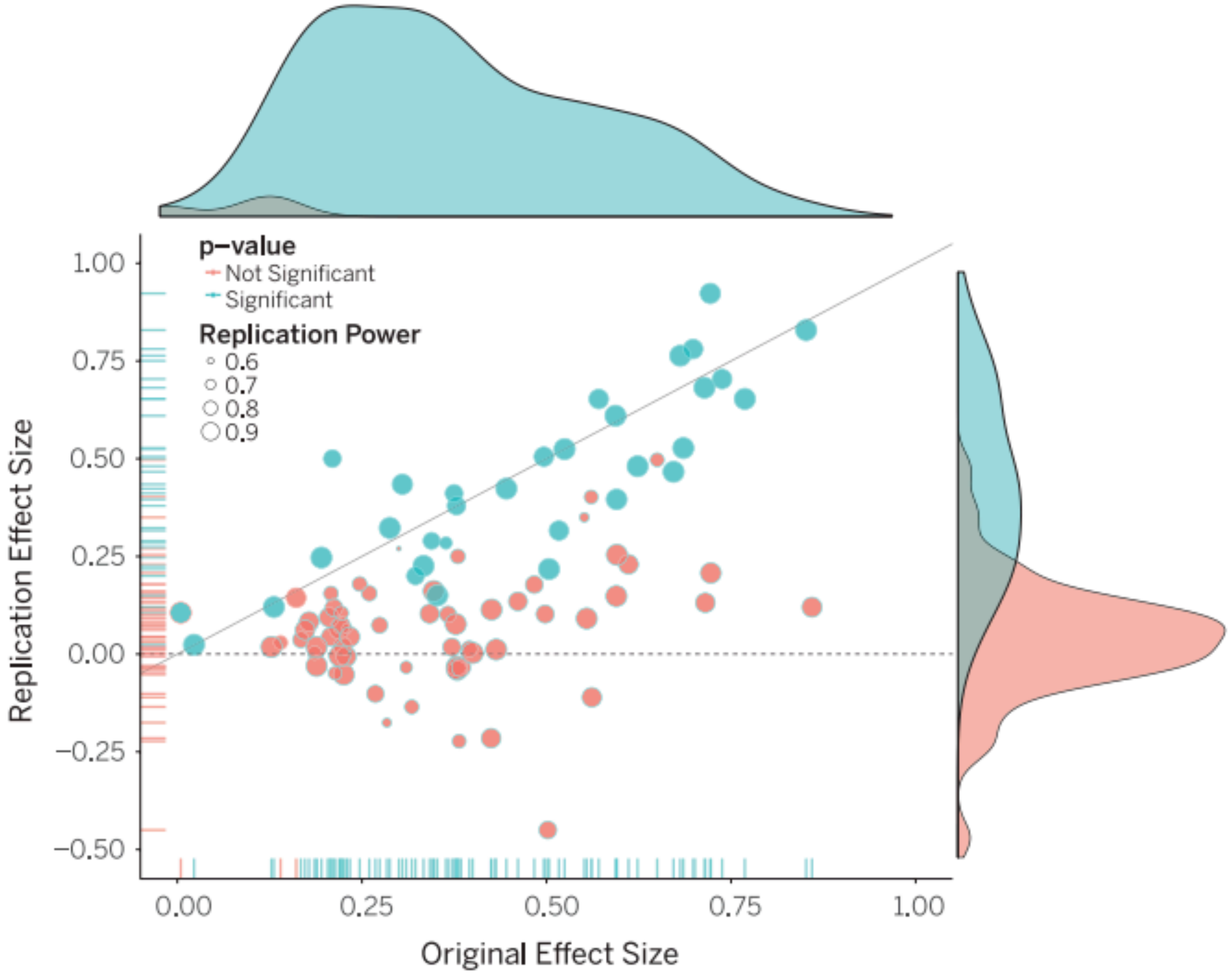
RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

(Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).)



(Goodman et al. 2016)

Types of reproducibility

Problem: Uncontrolled sources of variability can lead to dramatically different finding, even when using identical methods.

Results Reproducibility:

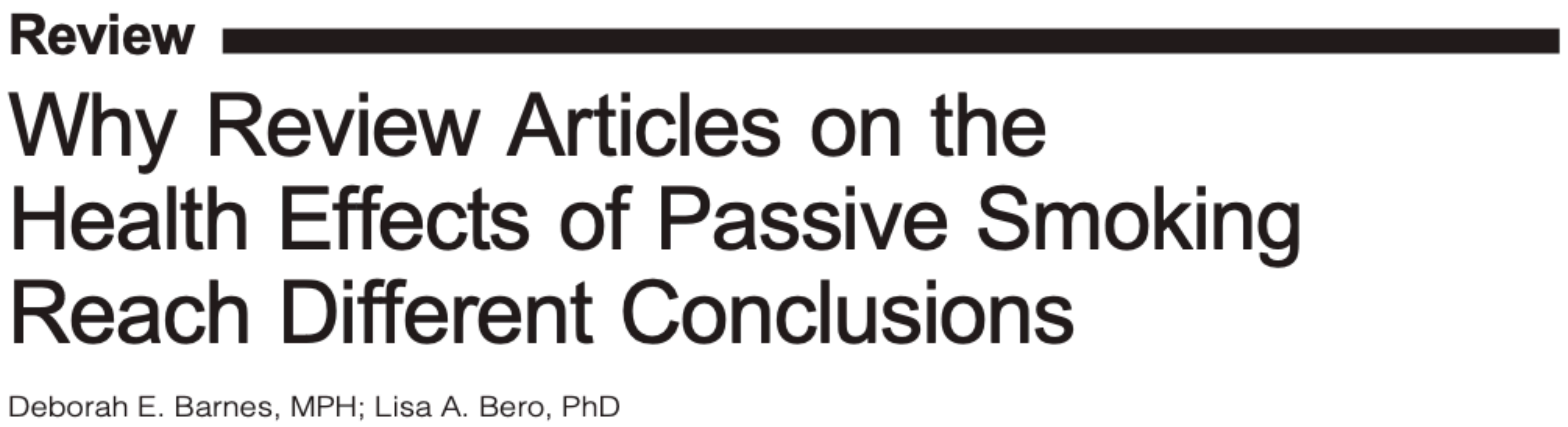
Obtaining the same results from independent experiments whose procedures are as closely matched as possible.

Solution: Standardization of data types, architectures, & quality assessments for sharing of data across research teams.

(Goodman et al. 2016)

Types of reproducibility

Problem: Variability across studies will lead to different conclusions from the same set of findings.



Review

Why Review Articles on the Health Effects of Passive Smoking Reach Different Conclusions

Deborah E. Barnes, MPH; Lisa A. Bero, PhD

(Barnes, D. E., & Bero, L. A. (1998). Why review articles on the health effects of passive smoking reach different conclusions. *JAMA*, *279*(19), 1566-1570.)

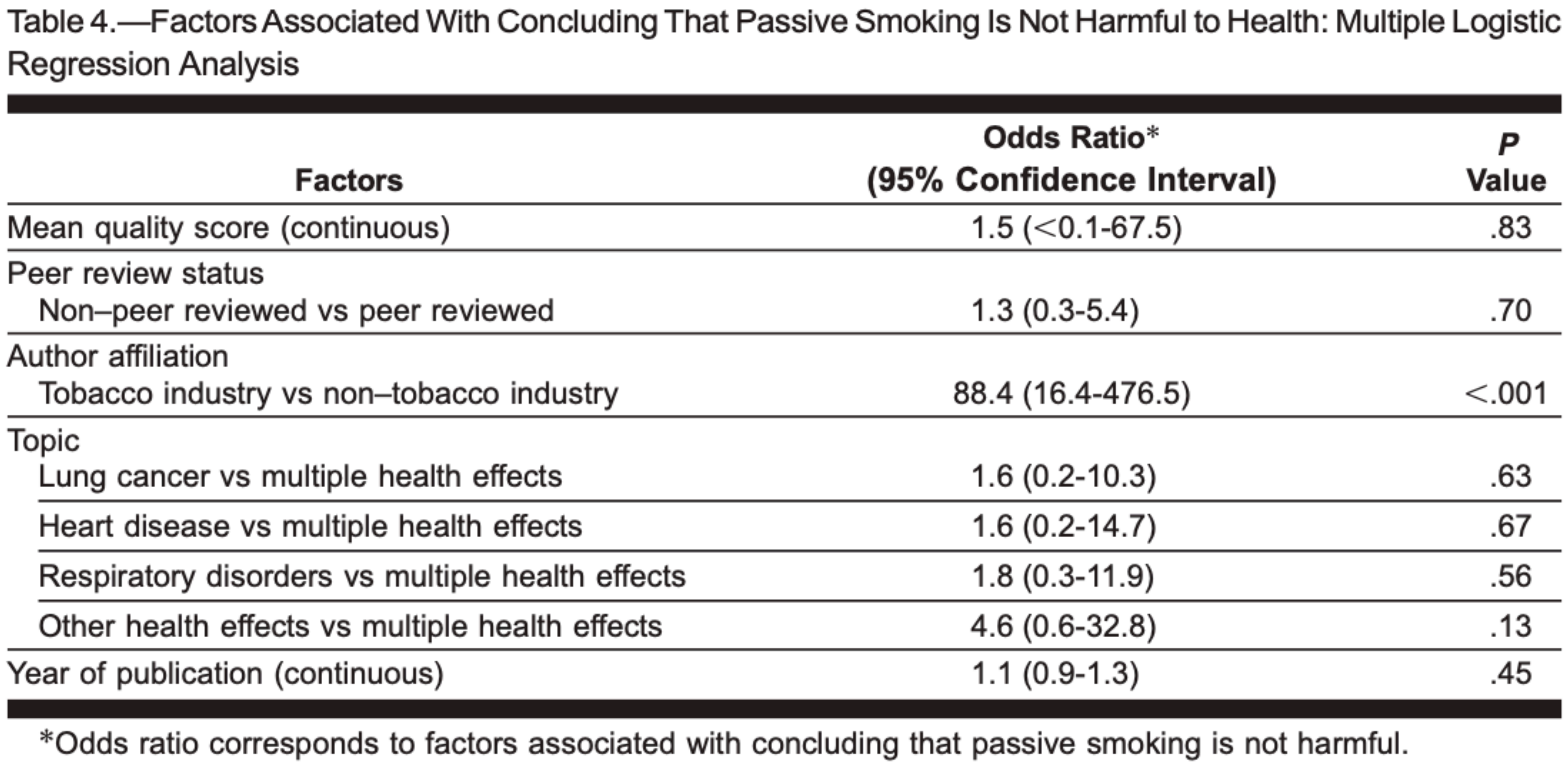


Table 4.—Factors Associated With Concluding That Passive Smoking Is Not Harmful to Health: Multiple Logistic Regression Analysis

Factors	Odds Ratio* (95% Confidence Interval)	*P* Value
Mean quality score (continuous)	1.5 (<0.1-67.5)	.83
Peer review status		
Non–peer reviewed vs peer reviewed	1.3 (0.3-5.4)	.70
Author affiliation		
Tobacco industry vs non–tobacco industry	88.4 (16.4-476.5)	<.001
Topic		
Lung cancer vs multiple health effects	1.6 (0.2-10.3)	.63
Heart disease vs multiple health effects	1.6 (0.2-14.7)	.67
Respiratory disorders vs multiple health effects	1.8 (0.3-11.9)	.56
Other health effects vs multiple health effects	4.6 (0.6-32.8)	.13
Year of publication (continuous)	1.1 (0.9-1.3)	.45

*Odds ratio corresponds to factors associated with concluding that passive smoking is not harmful.

(Goodman et al. 2016)

Types of reproducibility

Problem: Variability across studies will lead to different conclusions from the same set of findings.

Inferential Reproducibility:

Deriving the qualitatively similar conclusions from a set of similar studies or a replication/re-analysis of the same study.

Solution: Increased use of meta-analyses & incorporation of rigorous statistical analyses (including machine learning).

(Goodman et al. 2016)

Open science

A problem of communication

The Past (~1600s to 1990s)

- Medium: written journal articles or in person conferences.
  - Space & access limited.
  - Verbal communication of methods & findings

The Present (1990s to now)

- Medium: written journal articles, in person/virtual conferences, recorded talks, blogs, markdown notebooks, online repositories,
  - Few space constraints.
  - Direct transfer of methods & data.

The Open Science Movement

Definition:

“The movement to make scientific research & its dissemination accessible to all levels of an inquiring society, amateur or professional.”
- Wikipedia (adapted from Woelfle et al. 2011)

- Increase access to the *process* & *products* of science.

Rules for reproducible data science

How to do reproducible data science

1. **For every result, keep track of how it was produced.**
    - Analysis workflows
    - Carefully specified pipelines
    - Shell scripts

2. **Avoid manual data manipulation steps.**
    - Script everything
    - Use standard functions
    - Extensive documentation of manual steps that cannot be avoided

(Sandve et al. 2013)

How to do reproducible data science

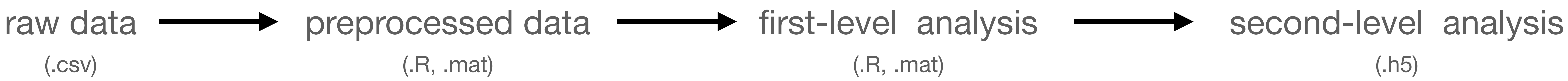
3. Archive the exact versions of all external programs used.

- Record version numbers of all programs used.
- Use dockers & containers.

4. Version control all custom scripts.

- Git, Subversion, etc.
- Well documented archive of scripts that are use.

5. Record all intermediary results, when possible in standardized formats.



raw data (.csv) → preprocessed data (.R, .mat) → first-level analysis (.R, .mat) → second-level analysis (.h5)

(Sandve et al. 2013)

How to do reproducible data science

6. **For analyses that include randomness, note the underlying random seed.**

- We cannot truly create randomness.
- All random number generators start with a “seed” number. Communicate that seed.

7. **Always store the raw data behind every plot.**

- Individual data tables for each plot.
- Standardized for many plotting tools.

8. **Generate hierarchical analysis outputs, allowing layers of increasing detail to be inspected.**

- Natural outcome of following rules #5 & #7.

(Sandve et al. 2013)

How to do reproducible data science

9. **Connect textual statements to underlying results.**
    - Value of markdown notebooks.
    - Show conclusions in the context of the data that leads to them.

10. **Provide public access to scripts, runs, & results.**
    - When possible, share everything on a public repository.
    - For proprietary data, share intermediary results.

(Sandve et al. 2013)

Ways to achieve the 10 rules for reproducible data science

- Use open source software (including interpreters): *R, python, Julia*

- Use flexible IDEs to develop and organize your code/scripts: *RStudio*

- Use markdown notebooks to *summarize* your investigation: *Jupyter, Rmarkdown*

- Use version control software: *Git, Subversion*

- Use public repositories for sharing data & code: *Github, Figshare, Kilthub*

Take home message

Adopting practices of the Open Science movement allows for achieving the goals of science in general & data science in particular.

Not: Science vs. Open Science

Is: Science vs. Closed Science