Reproducibility in Science

Thomas Nichols, PhD
Department of Statistics &
Warwick Manufacturing Group
University of Warwick

WDSI Summer School
Principles and Practice of Data Analysis for Reproducible Research in R

30 September 2016

Overview

- The Crises of Reproducibility
- TOP Principles

John Ioannidis' Crusade

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of

a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching

achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV.

- A careful argument for intense skepticism of modern scientific results
- Cited 3562 times (April 2016, Google Scholar)

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Study Positive Predictive Value

- Sampling Units
  - *Not* a set of subjects
  - A set of research hypotheses!
    - E.g. Hypothesis set in cognitive decline in aging:
      - Vitamin D reduces risk of cognitive decline
      - Exercise reduces risk of cognitive decline
      - Fish oil reduces risk of cognitive decline
      - …
- For a randomly selected study:
  - Given the study is positive, what is the probability the studied hypothesis is true?
  - I.e. what is the study PPV?

PPV Arithmetic

	True Hypothesis H+	False Hypothesis H-
Positive Finding D+	P(D+|H+) *Power* 1-β	P(D+|H-) *FPR* α
Negative Finding D-		
	P(H+)	P(H-)

- Notation
  - $R = N_T / N_F$ *odds of a true hypothesis*
    N_T = # true research hypotheses
    N_F = # false research hypotheses
  - P(H+) *probability of a true hypothesis*
  - Odds vs. probability
    - $P(H+) = N_T /(N_T + N_F) = R / (R+1)$

PPV Arithmetic

	True Hypothesis H+	False Hypothesis H-
Positive Finding D+	P(D+|H+) *Power* 1-β	P(D+|H-) *FPR* α
Negative Finding D-		
	P(H+)	P(H-)

- Bayes Theorem

$$P(H+) = R / (R+1)$$
$$P(H-) = 1 / (R+1)$$

$$PPV = P(H+|D+) = \frac{P(D+|H+)\,P(H+)}{P(D+|H+)\,P(H+) + P(D+|H+)\,P(H+)}$$

$$= \frac{(1-\beta)\,R / (R+1)}{(1-\beta)\,R / (R+1) + \alpha / (R+1)}$$

$$= \frac{(1-\beta)\,R}{(1-\beta)\,R + \alpha}$$

- PPV depends on power (1-β), odds of a true hypothesis (R) & false positive rate (FPR, α)

PPV Arithmetic

- When is PPV > ½?

$$0.5 > PPV = \frac{(1-\beta)\,R}{(1-\beta)\,R + \alpha} \quad \rightarrow \quad (1-\beta)R > \alpha$$

  - Note, (1-β)>α always true for a "unbiased" test
  - If R=1, PPV > ½
  - If R < ½, then PPV might < ½

- PPV & Power

$$PPV = \frac{(1-\beta)\,R}{(1-\beta)\,R + \alpha} = (1-\beta)\,\frac{R}{R + \alpha/(1-\beta)} \approx (1-\beta)$$

  - Lower the PPV, the lower the power

PPV Arithmetic

- PPV & "bias"
  - Suppose fraction *u* of all studies shouldn't have been published but are
    - i.e. won't have been published if no bias
    - Due to "vibration effects"
    - *Not* the α fraction of chance false positive studies
    - *Not* usual estimation bias per se
  - Then…

$$PPV = \frac{(1-\beta)\,R + u\,\beta\,R}{(1-\beta)\,R + u\,\beta\,R + \alpha + u(1-\alpha)}$$

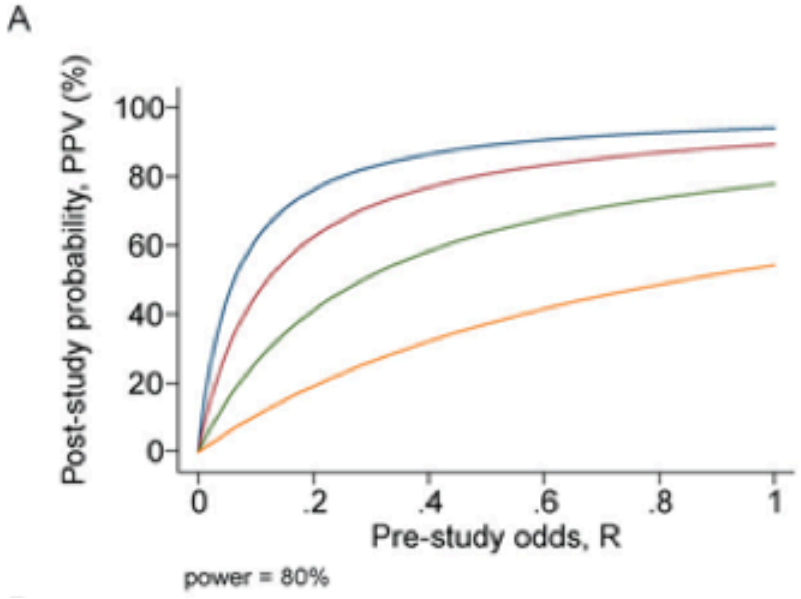
  - As *u* increases, PPV drops

Exploring study PPV

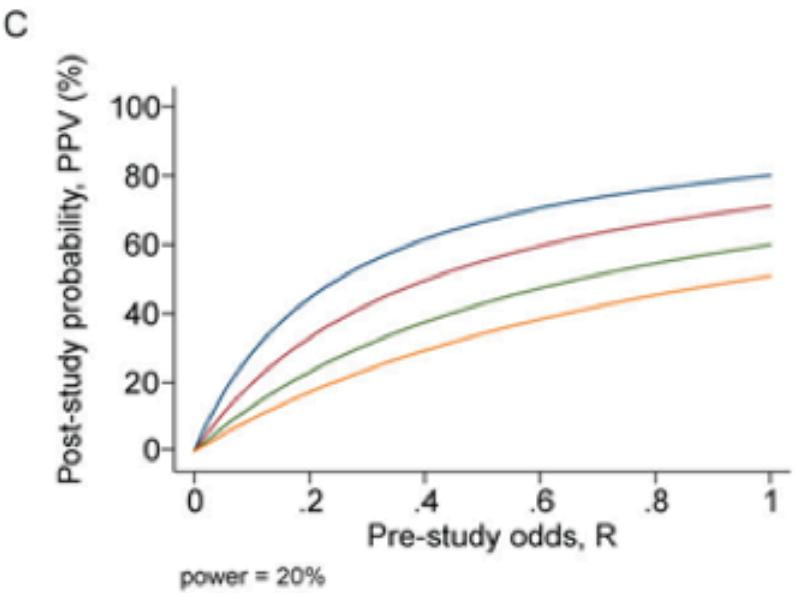
- PPV depends on u & power
  - Skepticism of a discipline (high 'bias' frequency u) translates to lower PPV

PPV vs. R - For different levels of bias *u*

Power = 80%



Power = 20%



u=0.05 u=0.20 u=0.50 u=0.80

Exploring “any” PPV

- Suppose *n* research teams all study a hypothesis
- Define “D+” as one or more of those teams getting a finding
  - They ‘busier’ the discipline, the lower the PPV

PPV vs. R - For number of research teams

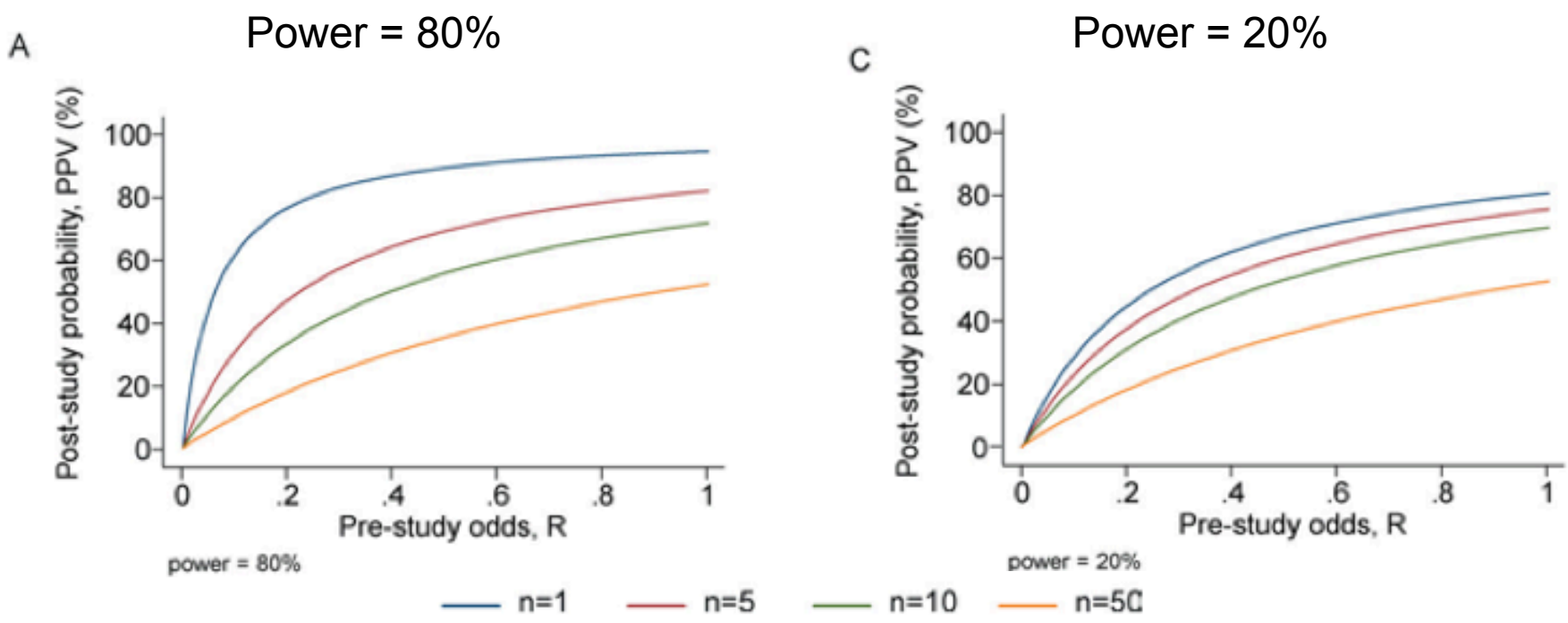


Table 4. PPV of Research Findings for Various Combinations of Power (1 – β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 – β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

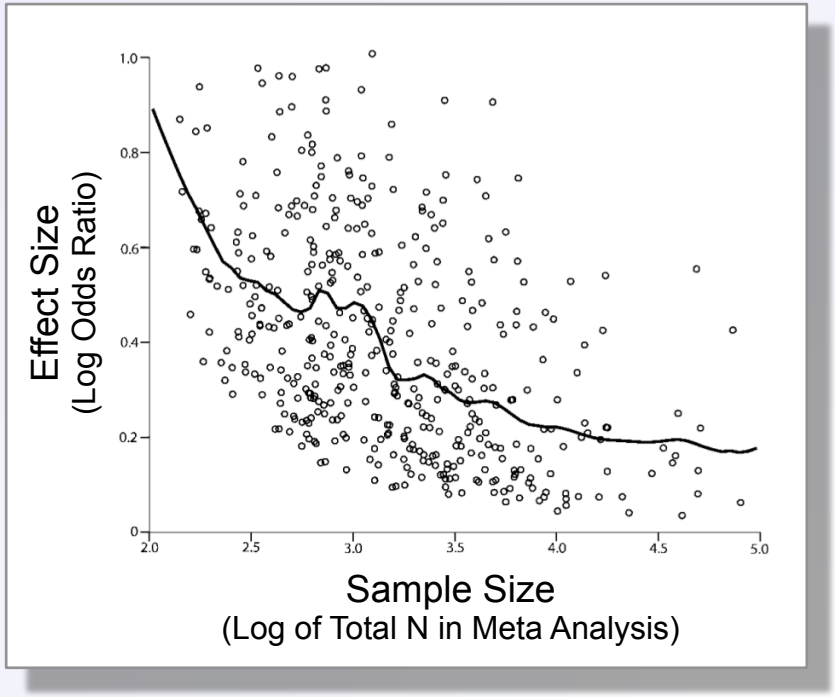
OK, but what's the evidence?

- This is a thought experiment
  - Sampling frame "Research hypotheses"
  - Many studies experience "bias", but this may take P-values from 0.0001 when then should be 0.005
- Is there really a problem here?
  - Canary in the coal mine, or
  - Chicken Little?

Exhibit A: Law of Small numbers

- Or “Winner’s Curse”
  - Small studies over-estimate effect size

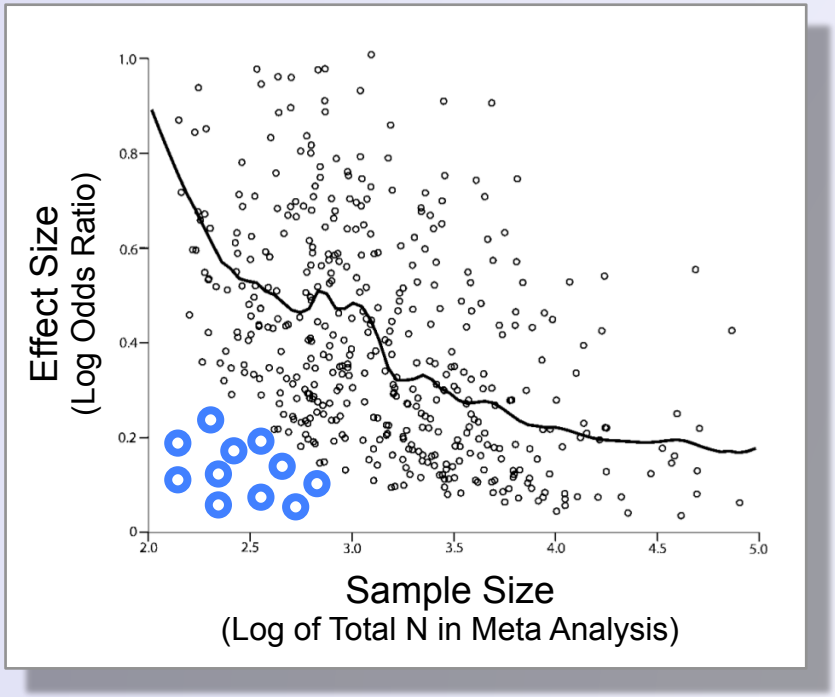
- ☐ 256 meta analyses for a dichotomous effect (odds ratio) from Cochrane database
- ☐ Studies with smallest N have biggest effect size!
  - ☑ Low N studies have low power
  - ☑ Low-power studies rarely succeed, but when they do, is result of randomly high effect or randomly small variance, biasing effect size
- ☐ Explains difficulty with replication



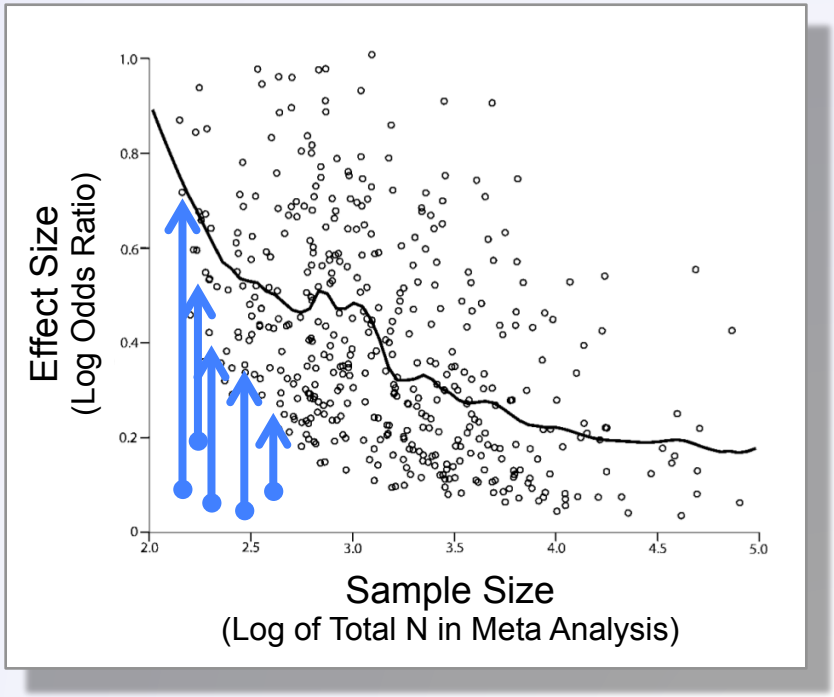
Ioannidis (2008). “Why most discovered true associations are inflated.” *Epidemiology*, 19(5), 640-8.

Two Problems

- Suppressed studies & Biased effects
  - P>0.05 not published
  - Biases that afflict small studies more than large studies



File drawer problem
(Unpublished non-significant studies)



Bias
(Fishing or Vibration Effects)

Vibration Effects

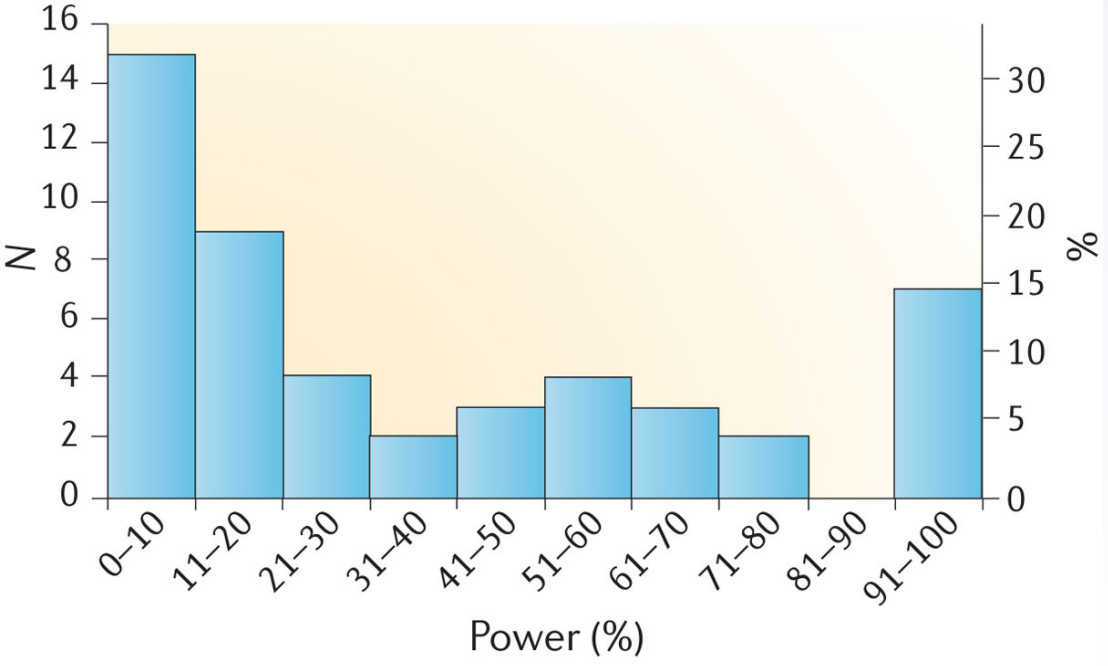
- Sloppy or nonexistent analysis protocols

> "Try voxel-wise whole brain, then cluster-wise, then if not getting good results, look for subjects with bad movement, if still nothing, maybe try a global signal regressor; if still nothing do SVC for frontal lobe, if not, then try DLPFC (probably only right side), if still nothing, will look in literature for xyz coordinates near my activation, use spherical SVC… surely that'll work!"

  - You stop when you get the result you expect
  - These "vibrations" can only lead to inflated false positives
- Afflicts well-intended researchers
  - Modern, "big data" scientific tools have multitude of preprocessing options, modeling choices
    - Pre-modelling normalisation options
    - Even more choices of options, covariates, interactions

Exhibit B: Studies chronically under powered

- Review of 730 neuroscience studies
  - Extracted from 48 meta analyses
  - Power of each of 730 studies calculated
- Median power **20%**
  - For 50% of studies, fewer than 1 in 5 replications will succeed!



Button et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews. Neuroscience, 14(5), 365–76.

Exhibit C: Mass replication

- Open Science Collaboration: Psychology
  - Replicated 100 new & classic studies
  - Effort of 270 scientists
- Each replication 'registered'
  - Carefully powered ($1-\beta \approx 90\%$)
  - Extensive peer review (usually with original authors contributing) in preparing study
  - Complete details of study protocol & analysis publically recorded and fixed

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. Science (New York, N.Y.), 349(6251), aac4716.

Exhibit C: Mass non-replication

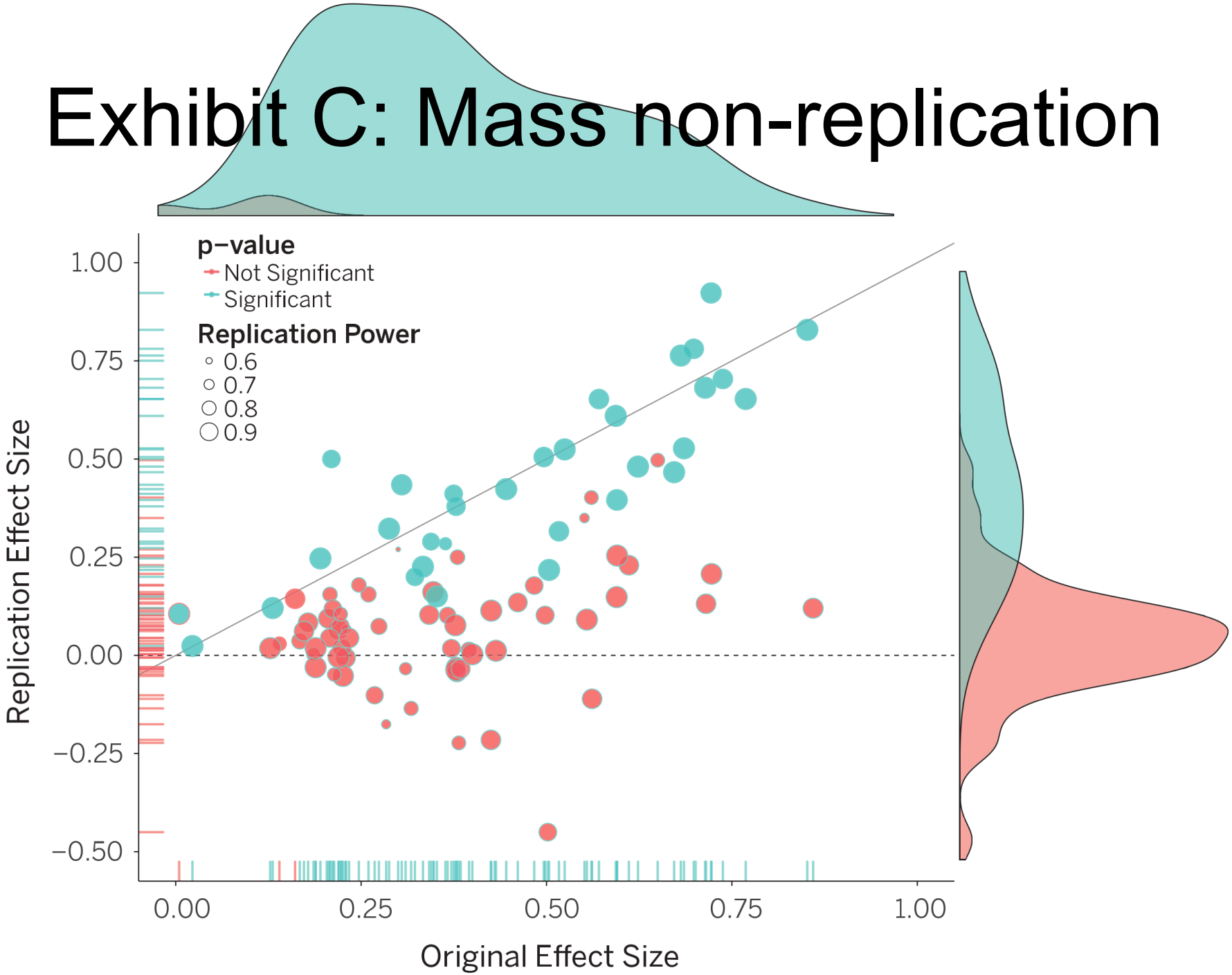
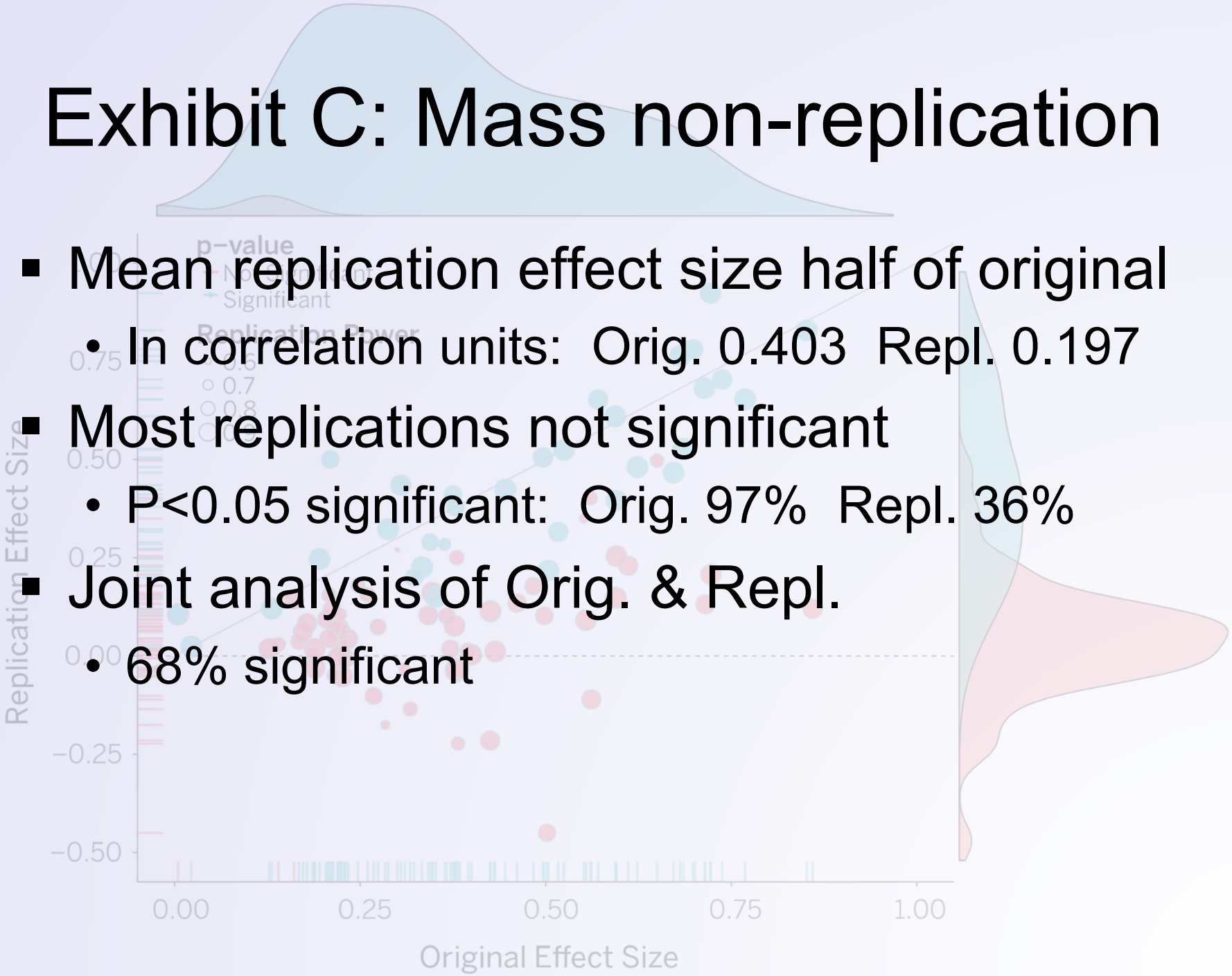


Exhibit C: Mass non-replication

- Mean replication effect size half of original
  - In correlation units: Orig. 0.403 Repl. 0.197
- Most replications not significant
  - P<0.05 significant: Orig. 97% Repl. 36%
- Joint analysis of Orig. & Repl.
  - 68% significant



What can be done?

- TOP – Transparency Openness Promotion
  - Advancing open science goals in service of reproducibly
  - Articulated by
    - Nosek et al. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*, 348(6242), 1422–5.
  - Provides 8 areas, 4 levels of success

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.

Elements of TOP

- Citation standards
- Data transparency
- Analytic methods (code) transparency
- Research materials transparency
- Design and analysis transparency
- Preregistration of studies
- Preregistration of analysis plans
- Replication

TOP Update (1/2)

- Citation standards
  - Citation of data, code and materials
  - Level 3: Complete citation of all data, code and materials
    - e.g. New *Science* standard
    - McNutt. (2016). Taking up TOP. Science, 352(6290), 1147–1147
- Data/Code/Materials transparency
  - Availability of data/code/materials
  - Level 3: Before pub., data, code & materials posted to trusted repository; reported analyses independently reproduced
    - e.g. “R” kite-mark in *Biostatistics*

TOP Uptake (2/2)

- Design and analysis transparency
  - Completely described design, following best practice
  - Level 3: Journal requires and enforces adherence to design standards for review and publication
    - Small steps: *Nature* / *Nature Neuroscience* check lists
- Preregistration of Study/Analysis Plan
  - Level 3: Required
- Replication
  - Facilitation of replication studies
  - Level 3: Registered report article type

Yes, the sky is falling.

- Many reasons to worry about validity of scientific literature
- Researchers need to…
  - Do power calculations
  - Disclose methods & findings transparently
  - Pre-register your study protocol and analysis plan
  - Make study materials and data available
  - Work collaboratively to increase power and replicate findings
    - Meta-Analyses

Open Discussion

- Has reliability/reproducibility of findings become an issue in your discipline? If so, how has the discpline reacted?
- What practices can you follow to ensure that someone else, given your data, could obtain your same results?
- What practices can you follow to ensure that someone else, starting from scratch, collecting new data, could obtain the same results that you have obtained?