香港中文大學
The Chinese University of Hong Kong

Power failure: Why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò

NATURE REVIEWS | NEUROSCIENCE VOLUME 14 | MAY 2013

Quentin Lam

October 3, 2024

Katherine Button

FOLLOW

GET MY OWN PROFILE

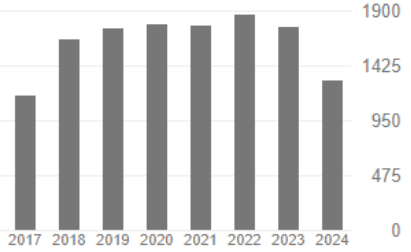
University of Bath
Verified email at bath.ac.uk - Homepage

Mental Health Disorders Anxiety Depression Cognition Scientific Rigor

TITLE	CITED BY	YEAR
Power failure: why small sample size undermines the reliability of neuroscience KS Button, JPA Ioannidis, C Mokrysz, BA Nosek, J Flint, ESJ Robinson, ... Nature reviews neuroscience 14 (5), 365-376	8479	2013
A manifesto for reproducible science MR Munafò, BA Nosek, DVM Bishop, KS Button, CD Chambers, ... Nature human behaviour 1 (1), 1-9	3042	2017
Deep impact: unintended consequences of journal rank B Brembs, K Button, M Munafò Frontiers in human Neuroscience 7, 45406	521	2013
Percie du Sert MR Munafò, BA Nosek, DVM Bishop, KS Button, CD Chambers N., Simonsohn, U., Wagenmakers, EJ, Ware, JJ, & Ioannidis, JPA, 1-9	360	2017
Minimal clinically important difference on the Beck Depression Inventory-II according to the patient's perspective KS Button, D Kounali, L Thomas, NJ Wiles, TJ Peters, NJ Welton, AE Ades, ... Psychological medicine 45 (15), 3269-3279	304	2015
Peer victimization during adolescence and risk for anxiety disorders in adulthood: a prospective cohort study LA Stapinski, L Bowes, D Wolke, RM Pearson, L Mahedy, KS Button, ... Depression and anxiety 31 (7), 574-582	264	2014

Cited by VIEW ALL

	All	Since 2019
Citations	15248	10287
h-index	30	27
i10-index	57	51



Public access VIEW ALL

2 articles 47 articles

not available available

Based on funding mandates

Professor Marcus Munafo

M.A.(Oxon.), M.Sc., Ph.D.(Soton.)

Expertise

I am interested in the relationship between health behaviours such as tobacco and alcohol use, and both physical and mental health outcome.

Current positions

Associate Pro Vice-Chancellor - Research Culture
Senior Team

Professor of Biological Psychology and MRC Investigator
School of Psychological Science

Contact

Marcus.Munafo@bristol.ac.uk

The Priory Road Complex
Priory Road
Clifton
Bristol, BS8 1TU

+44 117 455 8044

Press and media

Many of our academics speak to the media as experts in their field of research. If you are a journalist, please contact the University's Media and PR Team:

+44 117 428 2489

press-office@bristol.ac.uk

Edit profile

Introduction: What makes studies unreproducible

How low power affects study reliability without other biases

How low power affects study reliability **with other biases**

Method

Results

Discussions

Conclusions and future directions

Take home messages

Introduction: What makes studies unreproducible

Reproducibility crisis in psychology

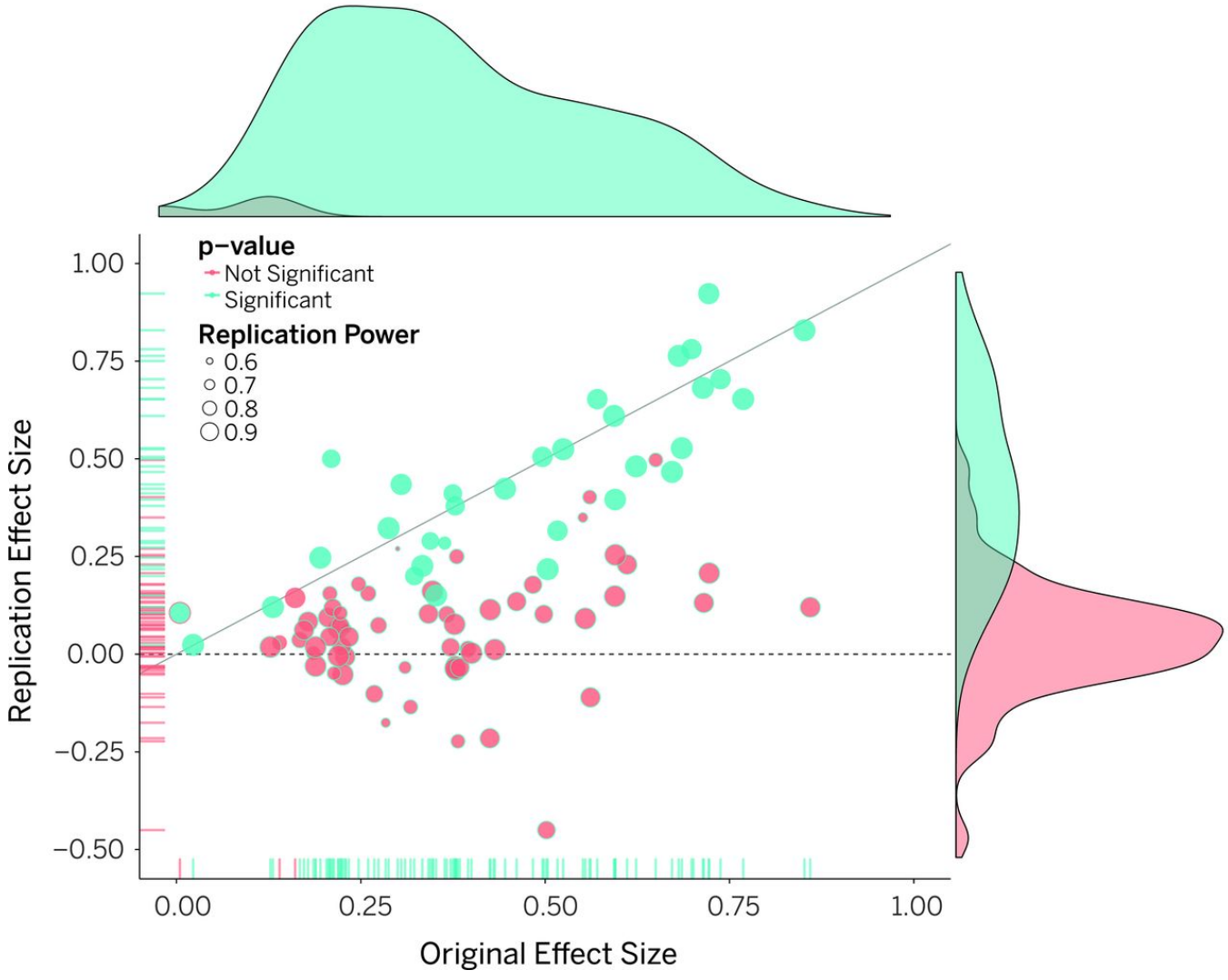


Figure 1: Original study effect size versus replication effect size (correlation coefficients) (Open Science Collaboration, 2015).

Statistical trap

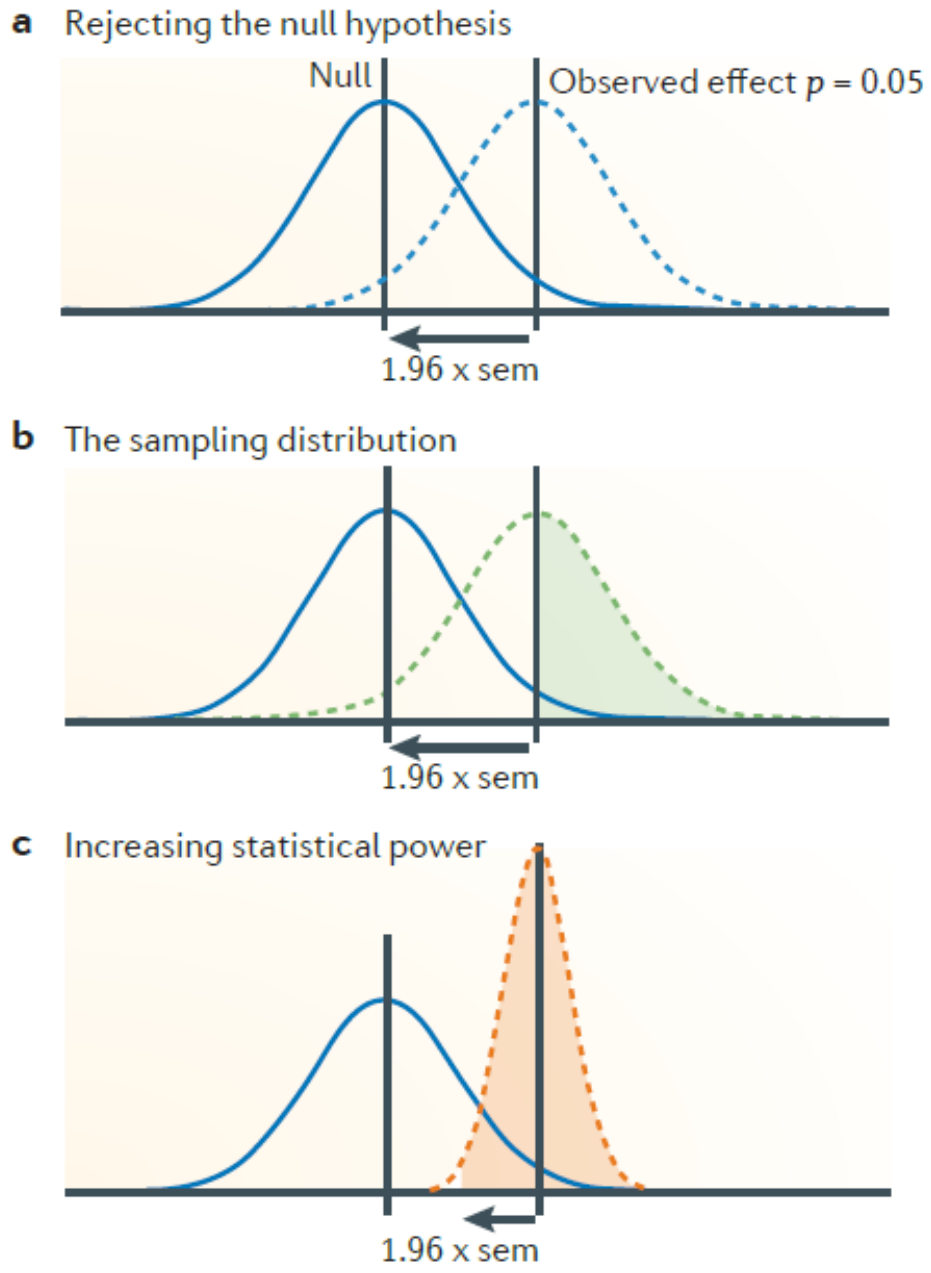
- Flexible experiment design
- Flexible statistical analyses
- Small sample size

$\Longrightarrow$

- Lower statistical power
- Higher false positive rate

In bio medicine, 97% genetic association studies have at least one false positive results (Sullivan, 2007).

Intuition of power



Power definition

The statistical power of a study is the probability of correctly rejecting the null hypothesis and detecting a statistically significant result (not committing a type II error) (Wheeler et al., 2014).

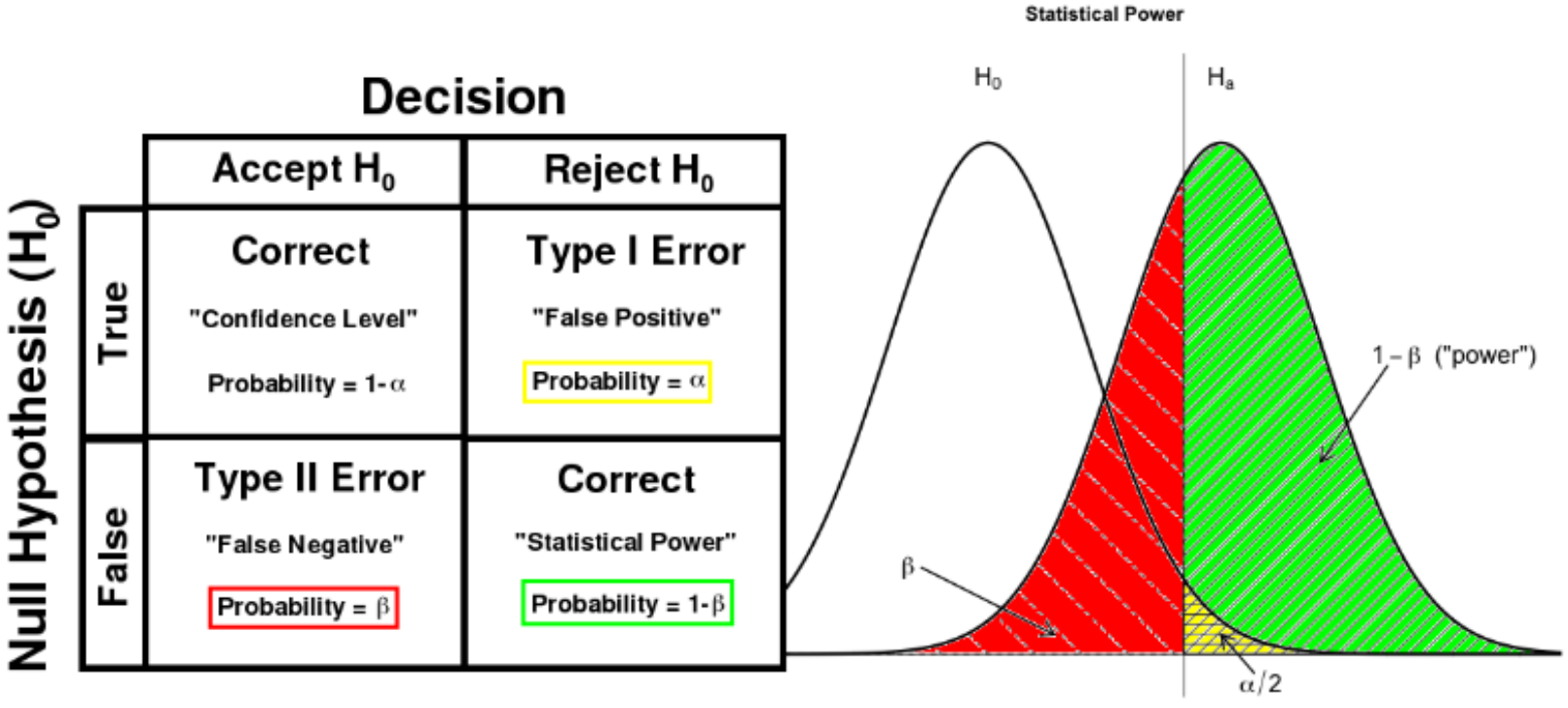


Figure 1: Left: Definitions of terminologies in a statistical test. Right: An illustration of power and significance level in a simple statistical test, where the left and right bell curves are the densities of the test statistics under the null hypothesis and the alternative hypothesis, respectively.[14]

Power examination

https://rpsychologist.com/d3/nhst/

How low power affects study reliability without other biases

Low power without other biases

3 problems in low power studies:

- Low probability of finding true effects
- Low positive predictive value when an effect is claimed
- Exaggerated estimate of the effect magnitude when a true effect is discovered

P 1/3: Low probability of finding true effects

Low power means that the chance of discovering effects that are genuinely true is low.

E.g., When studies in a given field are designed with a power of 20%, it means that if there are 100 genuine non-null effects to be discovered in that field, these studies are expected to discover only 20 of them.

P 2/3: Low positive predictive value when an effect is claimed

The lower the power of a study, the lower the probability that an observed effect that passes the required threshold of claiming its discovery (that is, reaching nominal statistical significance, such as p< 0.05) actually reflects a true effect.

$$\mathrm{PositivePredictiveValue(PPV)} = \frac{([1-\beta] \times \mathrm{R})}{([1-\beta] \times \mathrm{R} + \alpha)}$$

$$\Downarrow$$

$$\mathrm{PositivePredictiveValue(PPV)} = \frac{(\mathrm{Power} \times \mathrm{PrestudyOdds})}{(\mathrm{Power} \times \mathrm{PrestudyOdds} + \mathrm{TypeIErrorRate})}$$

R is the pre-study odds (that is, the odds that a probed effect is indeed non-null among the effects being probed).

P 2/3: Low positive predictive value when an effect is claimed

$$\text{PositivePredictiveValue(PPV)} = \frac{(\text{Power} \times \text{PrestudyOdds})}{(\text{Power} \times \text{PrestudyOdds} + \text{TypeIErrorRate})}$$

Power ↑, PPV ↑. Type I error rate ↑, PPV ↓.

E.g., if $\frac{1}{5}$ effects we test are expected to be truly non-null (that is, $R = \frac{P(DiscoverEffect)}{P(DiscoverNoEffect)} = \frac{1}{5-1} = 0.25$),

and the discovered effect $p < .05$, if have 20% power, then

$$PPV = \frac{0.20 \times 0.25}{0.20 \times 0.25 + 0.05} = \frac{0.05}{0.10} = 0.50$$

that is, only half of the findings would be correct.

if power = 0.8, then

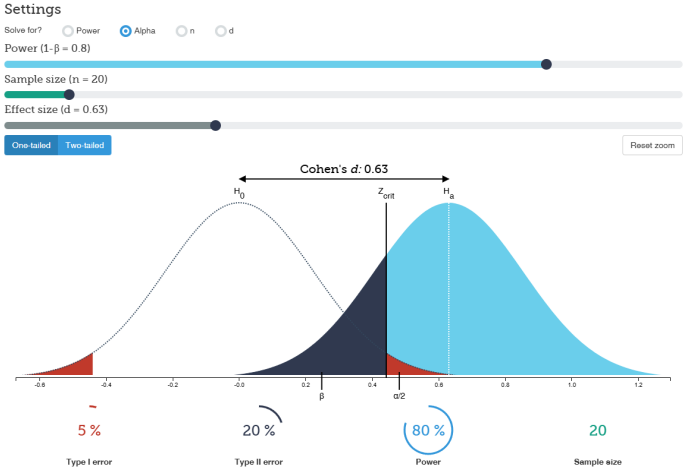
$$PPV = \frac{0.80 \times 0.25}{0.80 \times 0.25 + 0.05} = \frac{0.20}{0.25} = 0.80$$

which means 80% of our findings would be correct.

P 3/3: Exaggerated estimate of the effect magnitude when a true effect is discovered

Winner's curse: Lower power exaggerate the effect magnitude, especially based on the statistical significance(i.e., p value, BF value)

E.g., Medium True effect size, small studies can only pass the discovery threshold by overestimating the effect magnitude.



P 3/3: Exaggerated estimate of the effect magnitude when a true effect is discovered

To illustrate the winner's curse, suppose that true effect R(Pre-study odds) = 1.20.

Our study only has power that is 20% to detect an odds ratio of 1.20.

- Sampling variation
- Random error in the measurements

Because of random errors, our study may in fact find an odds ratio fluctuate around 1.20 (e.g., 1.00 or 1.60). Odds ratios of 1.00 or 1.20 will not reach statistical significance because of the small sample size. Only when the odds ratio is 1.60 can claim significance.

The winner's curse means that the 'lucky' scientist who makes the discovery in a small study is cursed by finding an inflated effect.

How low power affects study reliability with other biases

How low power affects study reliability with other biases

3 problems in low power studies:

- Vibration of effects
- Publication bias
- Low quality design

P 1/3: Vibration of effects

Vibration of effects: Study obtains different estimates of the effect magnitude depending on the analytical options it implements.

- Statistical model
- Definition of the variables of interest
- Use (or not) of adjustments for certain potential confounders but not others
- Filters to include or exclude specific observations

  .....

In small studies, the range of results that can be obtained owing to vibration of effects is wider than in larger studies, because the results are more uncertain and therefore fluctuate more in response to analytical changes.

P 2/3: Publication bias and selective reporting

Investigations into publication bias often examine whether small studies yield different results than larger ones.

A ‘negative’ result in a high-powered study cannot be explained away as being due to low power.

Thus reviewers and editors may be more willing to publish large studies, whereas they more easily reject a small ‘negative’ study as being inconclusive or uninformative.

The protocols of large studies are also more likely to be publicly available, so that deviations in the analysis plans and choice of outcomes may become obvious more easily.

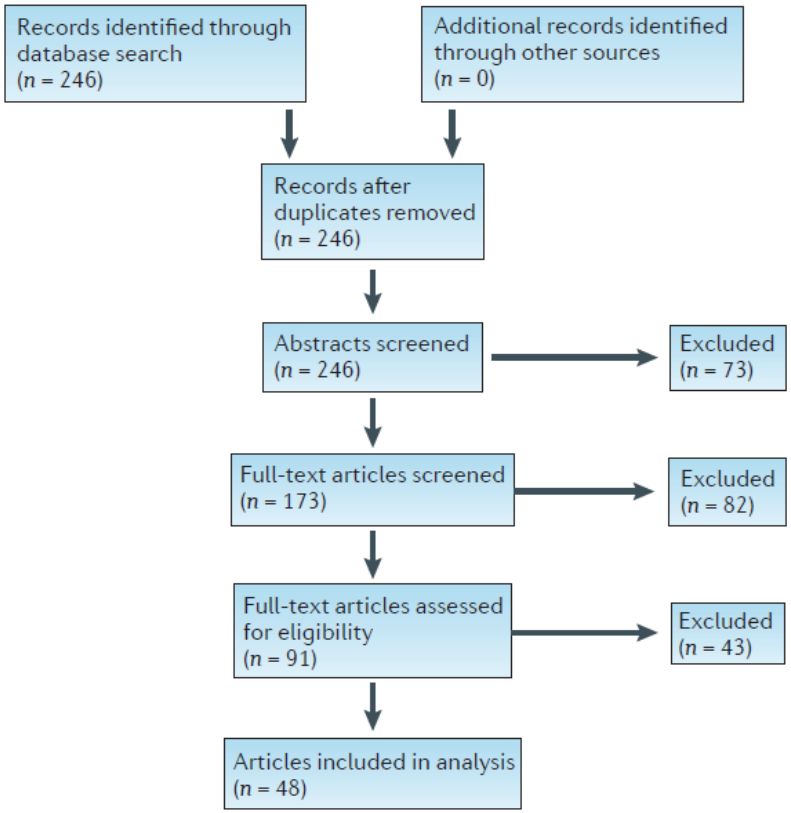
P 3/3: Low quality design

Large studies often require more funding and personnel resources, so the designs are examined more carefully before data collection, and analysis and reporting may be more structured.

A favour of small studies may occur if the small studies are meticulously designed and collect high-quality data (and therefore are forced to be small) and if large studies ignore or drop quality checks in an effort to include as large a sample as possible.

Method

Meta-analyses inclusion flow diagram



- Two authors (K.S.B. and M.R.M.) independently screened all papers
- Exclude no abstract
- Exclude no full-text
- Include 49 meta-analyses with 730 studies
- If meta-analyses have overlapped studies, include one with more studies overall
- If studies have missing data, then don't include it in the meta-analyses

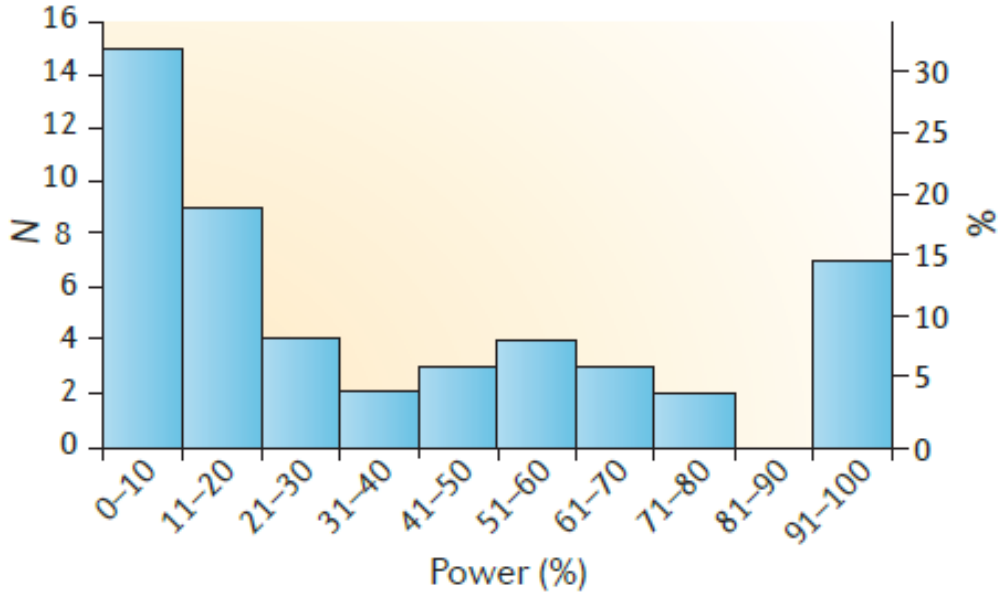
Results

Median power in Neuroscience

The median power is 21%

Test for an excess of statistical significance: 349 out of 740 studies significant was higher than the number expected (254; $p < 0.0001$).

Almost 50% of studies had an average power lower than 20%, median statistical power in reliable sample size studies is 18%



Excess significance bias in Neuroimaging and animal studies

the fMRI studies median statistical power was 8% across 461 individual studies contributing to 41 separate meta-analyses.

Table 2 | **Sample size required to detect sex differences in water maze and radial maze performance**

	Total animals used	Required *N* per study		Typical *N* per study		Detectable effect for typical *N*	
		80% power	95% power	Mean	Median	80% power	95% power
Water maze	420	134	220	22	20	$d = 1.26$	$d = 1.62$
Radial maze	514	68	112	24	20	$d = 1.20$	$d = 1.54$

Meta-analysis indicated an effect size of Cohen's $d = 0.49$ for water maze studies and $d = 0.69$ for radial maze studies.

Discussions

Implications for the likelihood that a research finding reflects a true effect

The average statistical power of studies in the field of neuroscience is probably no more than between ∼8% and ∼31%,

$$\text{PositivePredictiveValue(PPV)} = \frac{(\text{Power} \times \text{PrestudyOdds})}{(\text{Power} \times \text{PrestudyOdds} + \text{TypeIErrorRate})}$$

Power ↑, PPV ↑. Type I error rate ↑, PPV ↓.

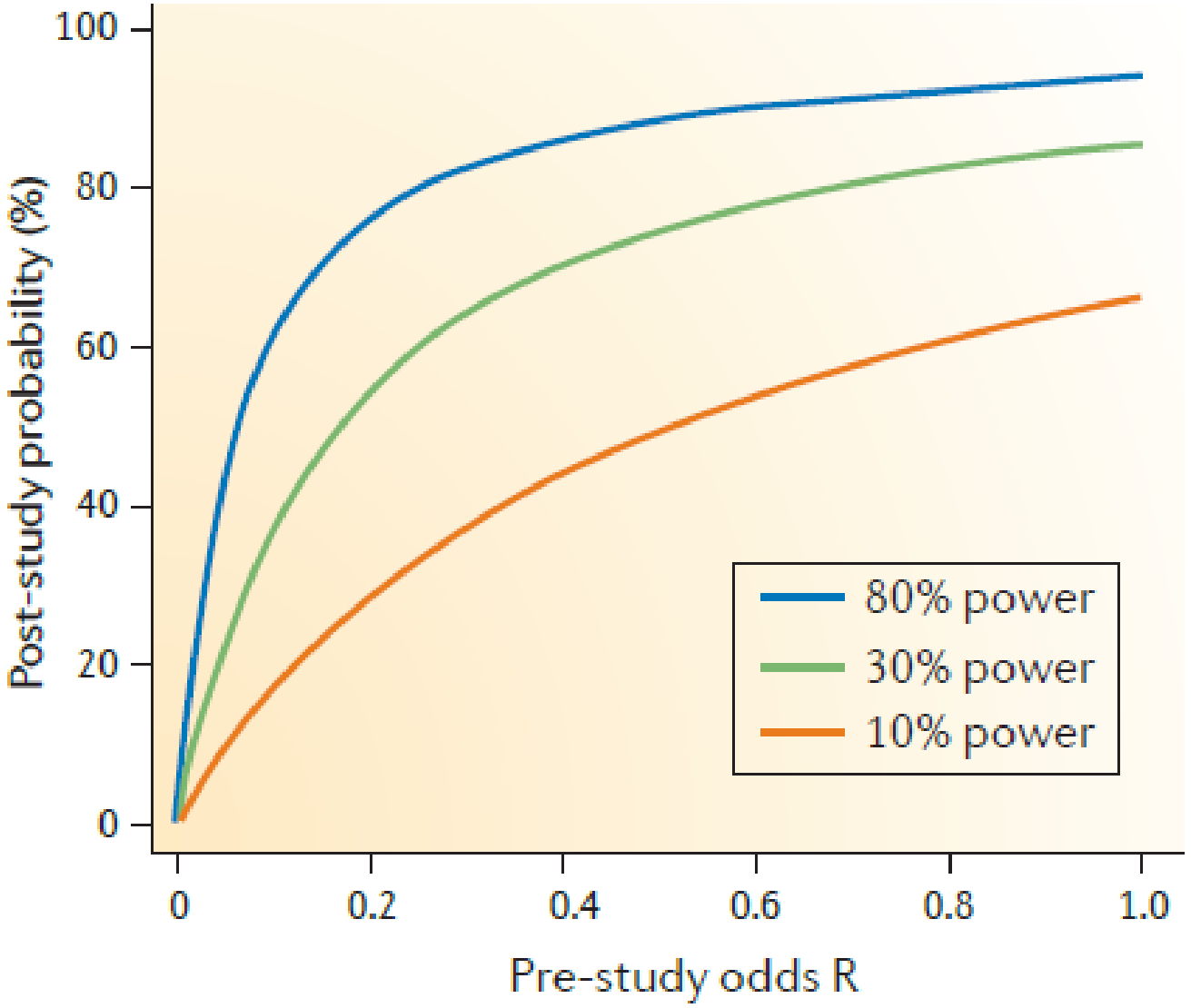
E.g., if $\frac{1}{5}$ effects we test are expected to be truly non-null (that is, $R = \frac{P(DiscoverEffect)}{P(DiscoverNoEffect)} = \frac{1}{5-1} = 0.25$),

and the discovered effect $p < .05$, if have 20% power, then

$$PPV = \frac{0.20 \times 0.25}{0.20 \times 0.25 + 0.05} = \frac{0.05}{0.10} = 0.50$$

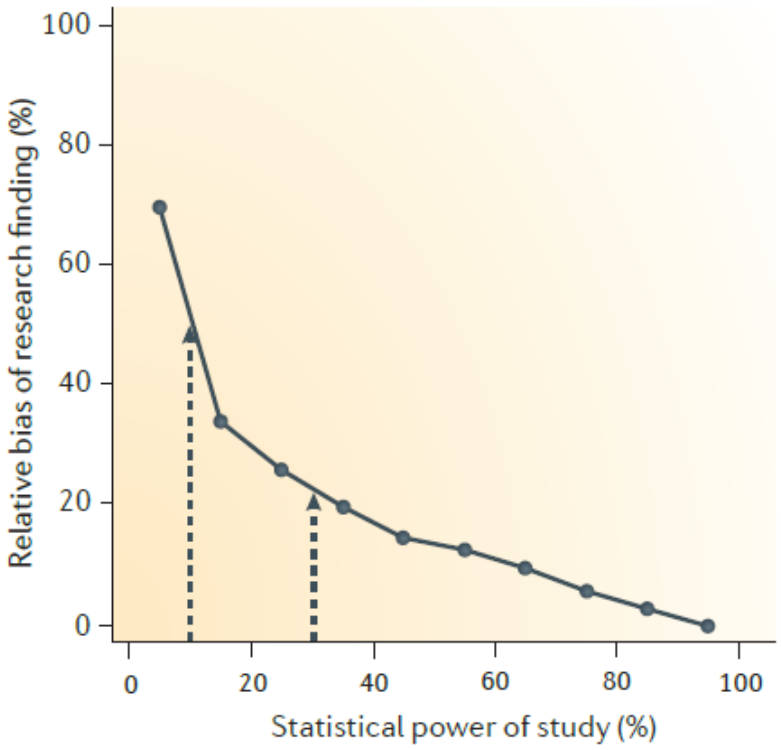
that is, only half of the findings would be correct.

Positive predictive value as a function of the pre-study odds of association for different levels of statistical power



The winner's curse: Effect size inflation as a function of statistical power

These simulations suggest that initial effect estimates from studies powered between ~8% and ~31% are likely to be inflated by 25% to 50%. Inflated effect estimates make it difficult to determine an adequate sample size for replication studies, increasing the probability of type *II* errors.



Real study power is even lower

Above simulation only consider R and power, did not consider several other biases that reduce the probability that a research finding reflects a true effect.

Excess of significance test showed these studies effect sizes are inflated.

Conclusions and future directions

Conclusions and future directions

Box 2 | **Recommendations for researchers**

Perform an a priori power calculation
Use the existing literature to estimate the size of effect you are looking for and design your study accordingly. If time or financial constraints mean your study is underpowered, make this clear and acknowledge this limitation (or limitations) in the interpretation of your results.

Disclose methods and findings transparently
If the intended analyses produce null findings and you move on to explore your data in other ways, say so. Null findings locked in file drawers bias the literature, whereas exploratory analyses are only useful and valid if you acknowledge the caveats and limitations.

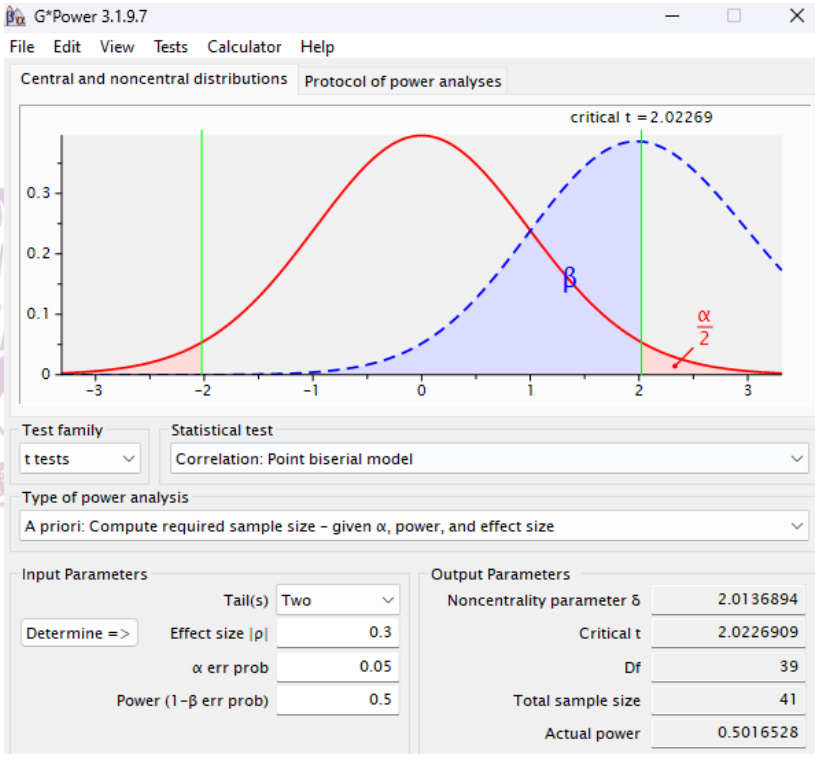
Pre-register your study protocol and analysis plan
Pre-registration clarifies whether analyses are confirmatory or exploratory, encourages well-powered studies and reduces opportunities for non-transparent data mining and selective reporting. Various mechanisms for this exist (for example, the Open Science Framework).

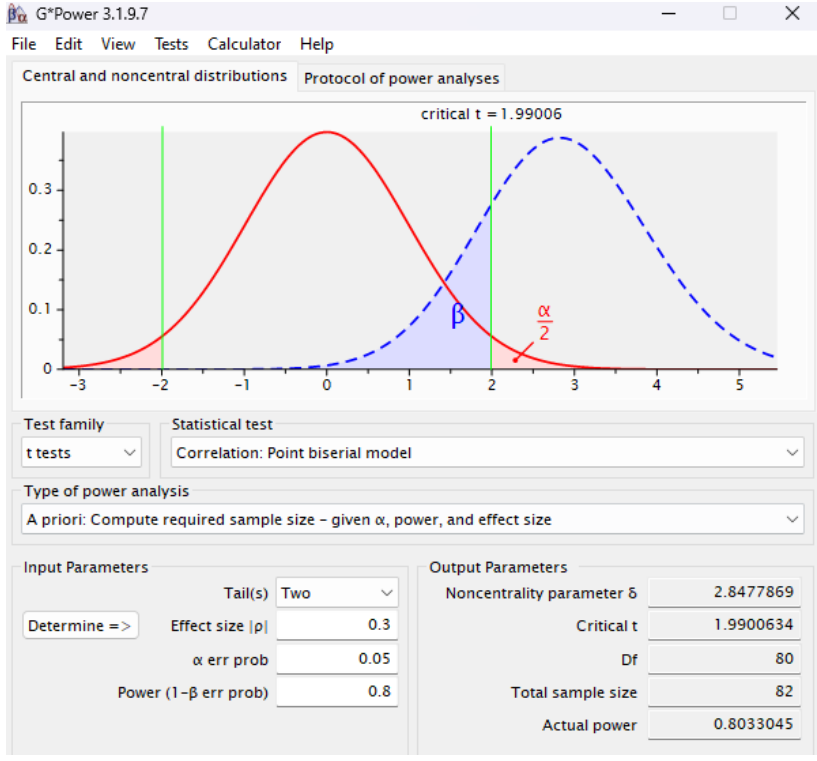
Make study materials and data available
Making research materials available will improve the quality of studies aimed at replicating and extending research findings. Making raw data available will enhance opportunities for data aggregation and meta-analysis, and allow external checking of analyses and results.

Work collaboratively to increase power and replicate findings
Combining data increases the total sample size (and therefore power) while minimizing the labour and resource impact on any one contributor. Large-scale collaborative consortia in fields such as human genetic epidemiology have transformed the reliability of findings in these fields.

Take home messages

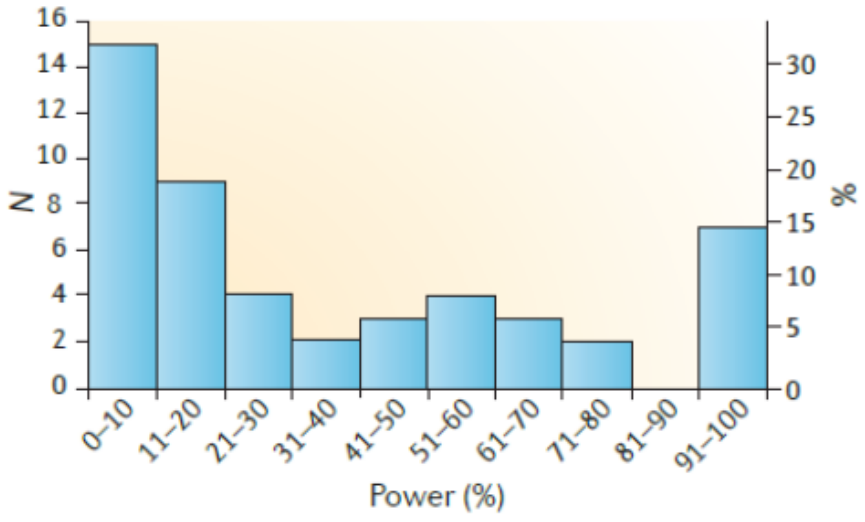
Sample size should at least double for replication study





Hints that sample size should be over 80

Almost 50% of studies had an average power lower than 20%, median statistical power in reliable sample size studies is 18%



These seven meta-analyses were all broadly neurological in focus and were based on relatively small contributing studies — four out of the seven meta-analyses did not include any study with over 80 participants.

Thank you for your attention!

References

Open Science Collaboration. (2015). **Estimating the reproducibility of psychological science.** Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Sullivan, P. F. (2007). **Spurious Genetic Associations.** Biological Psychiatry, 61(10), 1121–1126. https://doi.org/10.1016/j.biopsych.2006.11.010

Wheeler, N., Xu, Y., Gok, A., Kidd, I., Bruckman, L., Sun, J., & French, R. (2014, June 1). **Data Science Study Protocols for Investigating Lifetime and D**