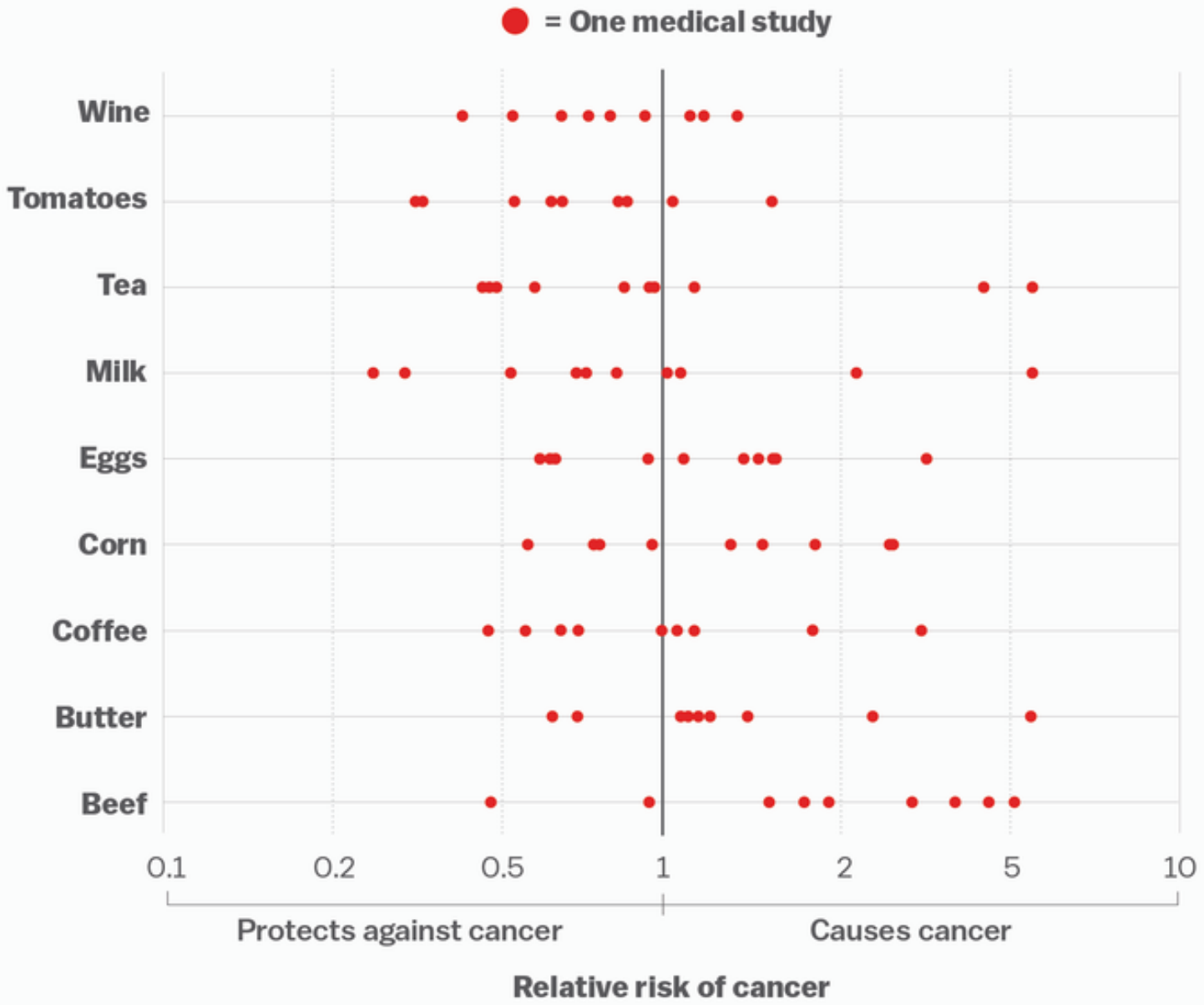
# Reproducibility, replicability, or communication crisis?

Valentin Amrhein

University of Basel

# Everything we eat both causes and prevents cancer




SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

Vox


Sarah2 /Shutterstock

# This is why you shouldn't believe that exciting new medical study

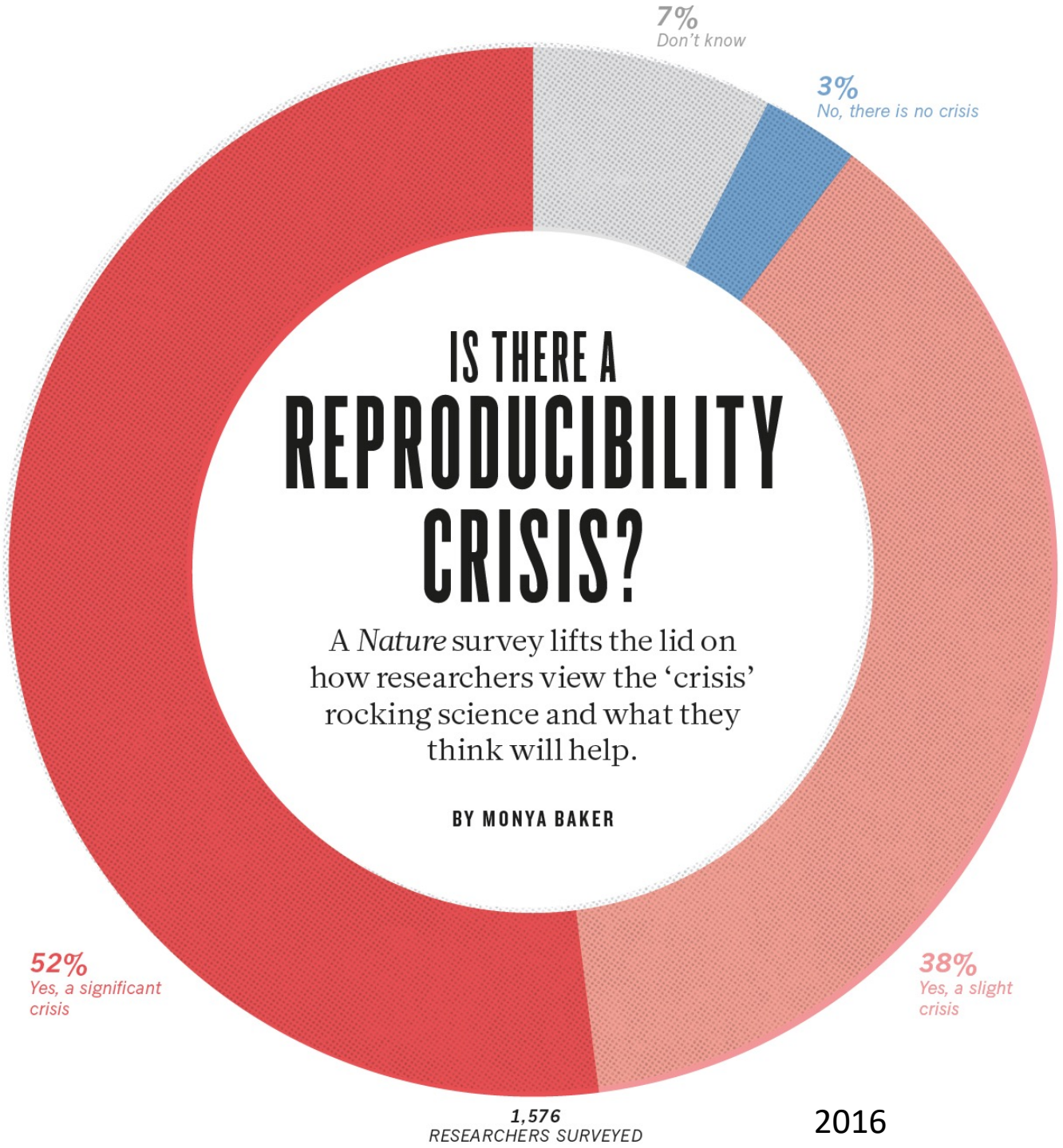
By Julia Belluz | @juliaoftoronto | Updated Feb 27, 2017, 9:18am EST



Unreliable research

# Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not


7%
Don't know
3%
No, there is no crisis
IS THERE A
REPRODUCIBILITY
CRISIS?
A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.
BY MONYA BAKER
52%
Yes, a significant crisis
38%
Yes, a slight crisis
1,576
RESEARCHERS SURVEYED

2016

What does research reproducibility mean?
Goodman, Fanelli, Ioannidis 2016, Science Translational Medicine

**Methods reproducibility**
Is enough detail provided in a paper/protocol so that the study procedures could be exactly repeated?

**Results reproducibility**, or **replicability**
Can we obtain the same results from an independent replication of a study?

**Inferential reproducibility**
Can we draw qualitatively similar conclusions from an independent replication of a study?

# New study says studies are wrong

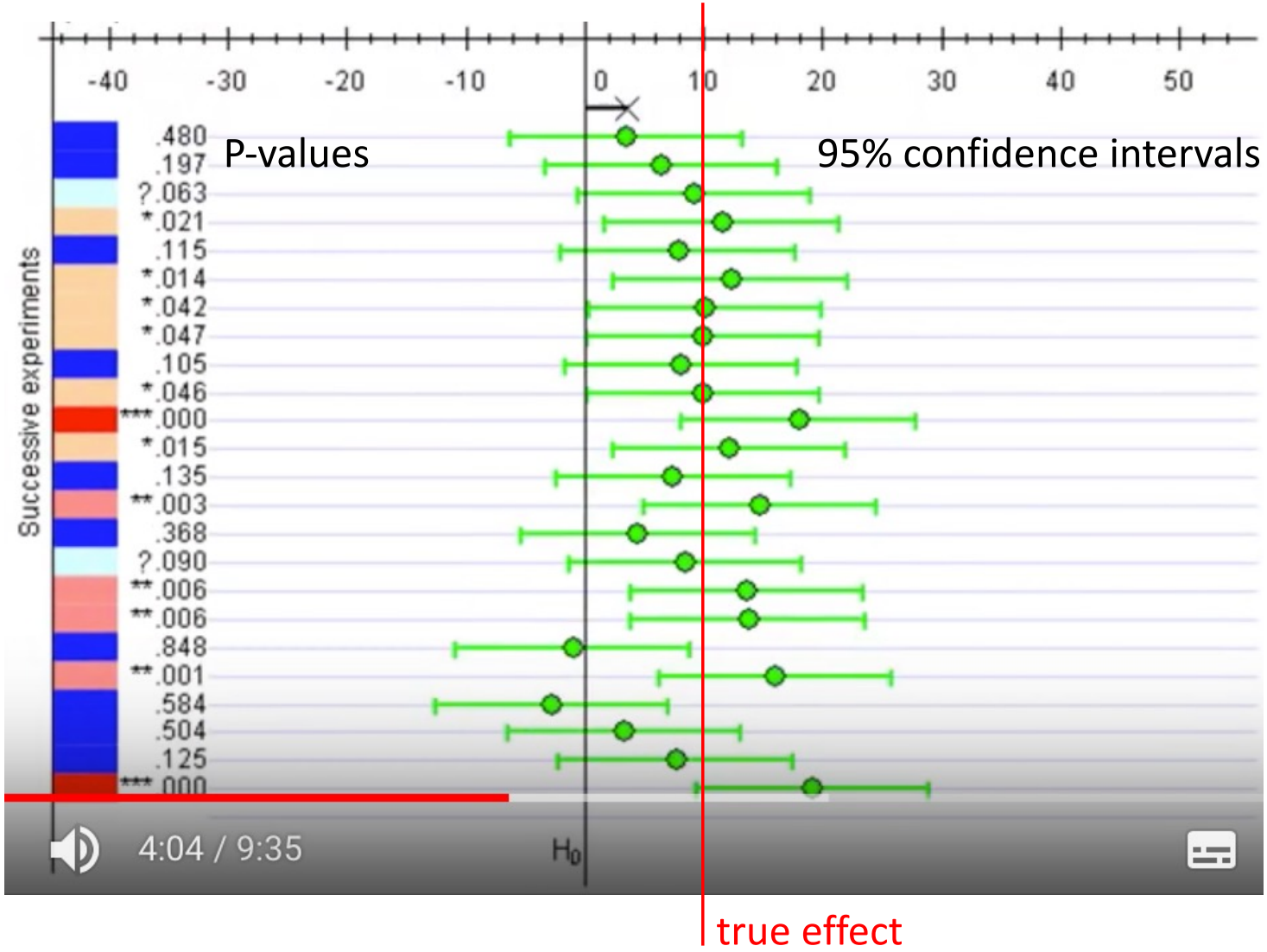
AFP RELAXNEWS | AUG 28, 2015 | 8:37 AM

Some studies aren't worth stressing over. (Halfpoint/shutterstock.com)

A team of 270 scientists tried reproducing 100 psychology and social science studies that had been published in three top peer-reviewed U.S. journals in 2008.

Just 39 percent came out with same results as the initial reports, said the findings in the journal Science.

# Geoff Cumming 2009: Dance of the P-values (youtube)




Simulated random samples from a population with true effect = 10, SD = 20, n = 32, power = 52%

**These results would usually not be considered "the same"**

# The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value.

## Don't blame the P-value

The P-value is not supposed to be 'reliable' in the sense of staying put. Its fickleness indicates variation in the data from sample to sample.

Just as effect size estimates vary among samples, P-values vary as well, because they are calculated from effect size estimates.

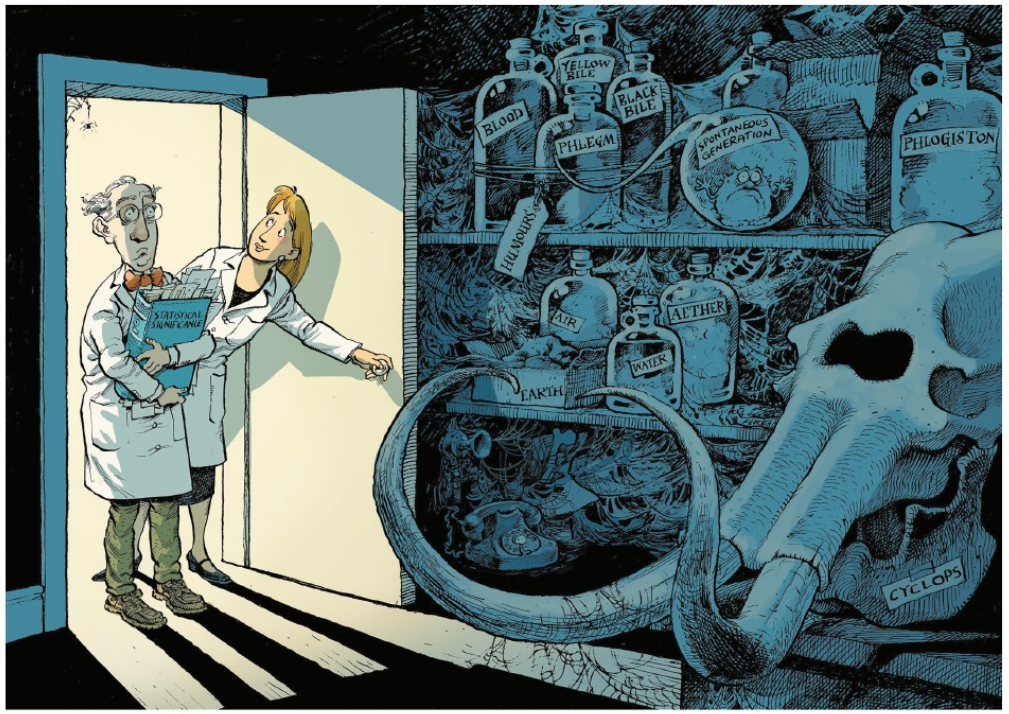
But making "yes" or "no" decisions based on P-value thresholds (**dichotomania**) from single studies means having **overconfidence**.


Amrhein, Trafimow, Greenland 2019, The American Statistician

# COMMENT



ILLUSTRATION BY DAVID PARKINS




# Retire statistical significance


**Valentin Amrhein, Sander Greenland, Blake McShane** and more than 800 signatories


call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

## PERVASIVE PROBLEM

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 ▶

"Mathematical vs. scientific significance"
Boring **1919**, Psychological Bulletin

"The fallacy of the null-hypothesis significance test"
Rozeboom **1960**, Psychological Bulletin

"The earth is round ($p < .05$)"
Cohen **1994**, American Psychologist

"The insignificance of statistical significance testing"
Johnson **1999**, Journal of Wildlife Management


**2016**
Statement on p-values by the American Statistical Association

**2019**
Special issue in The American Statistician with 43 papers on "Statistical inference in the 21st century: A world beyond $p < 0.05$"

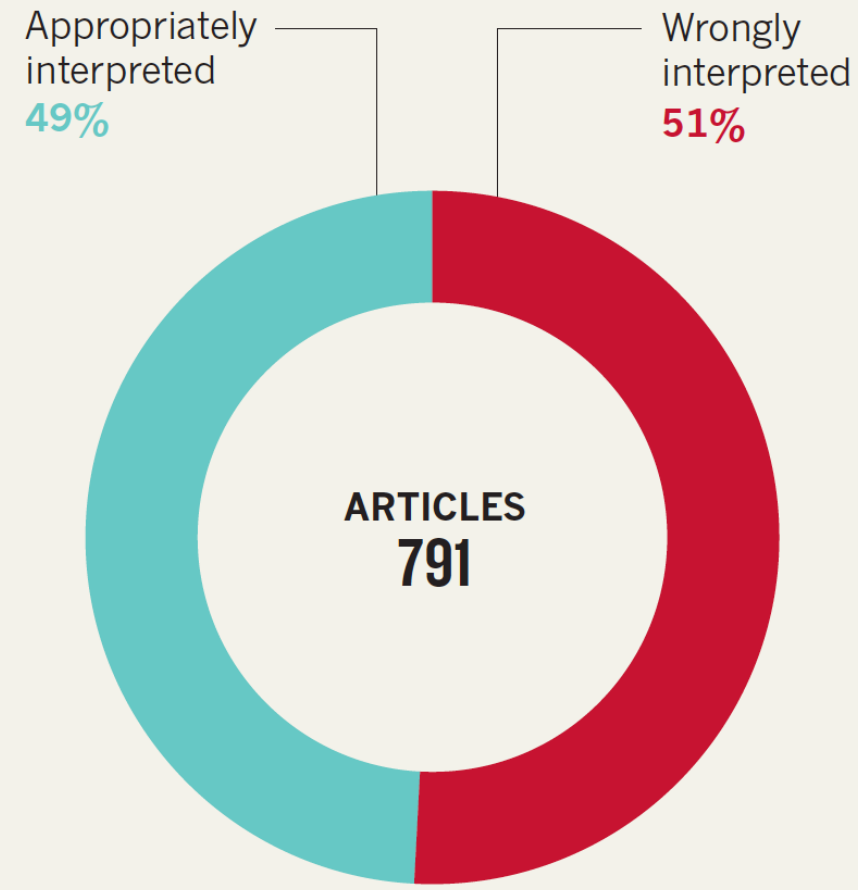
same-sex pairs (Tables 2, 3). While there was no sex difference in the mean duration of USVs (females: 99.71 $\pm$ 7.28 ms; males: 76.79 $\pm$ 9.18 ms; Mann–Whitney $U$ test: $U = 2271.5$, $N_1 = 114$, $N_2 = 48$, $P = 0.104$), frequency of USVs (females: 59.47 $\pm$ 0.90 kHz;

USVs = ultrasonic vocalizations

2015, Animal Behaviour

## WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.




*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

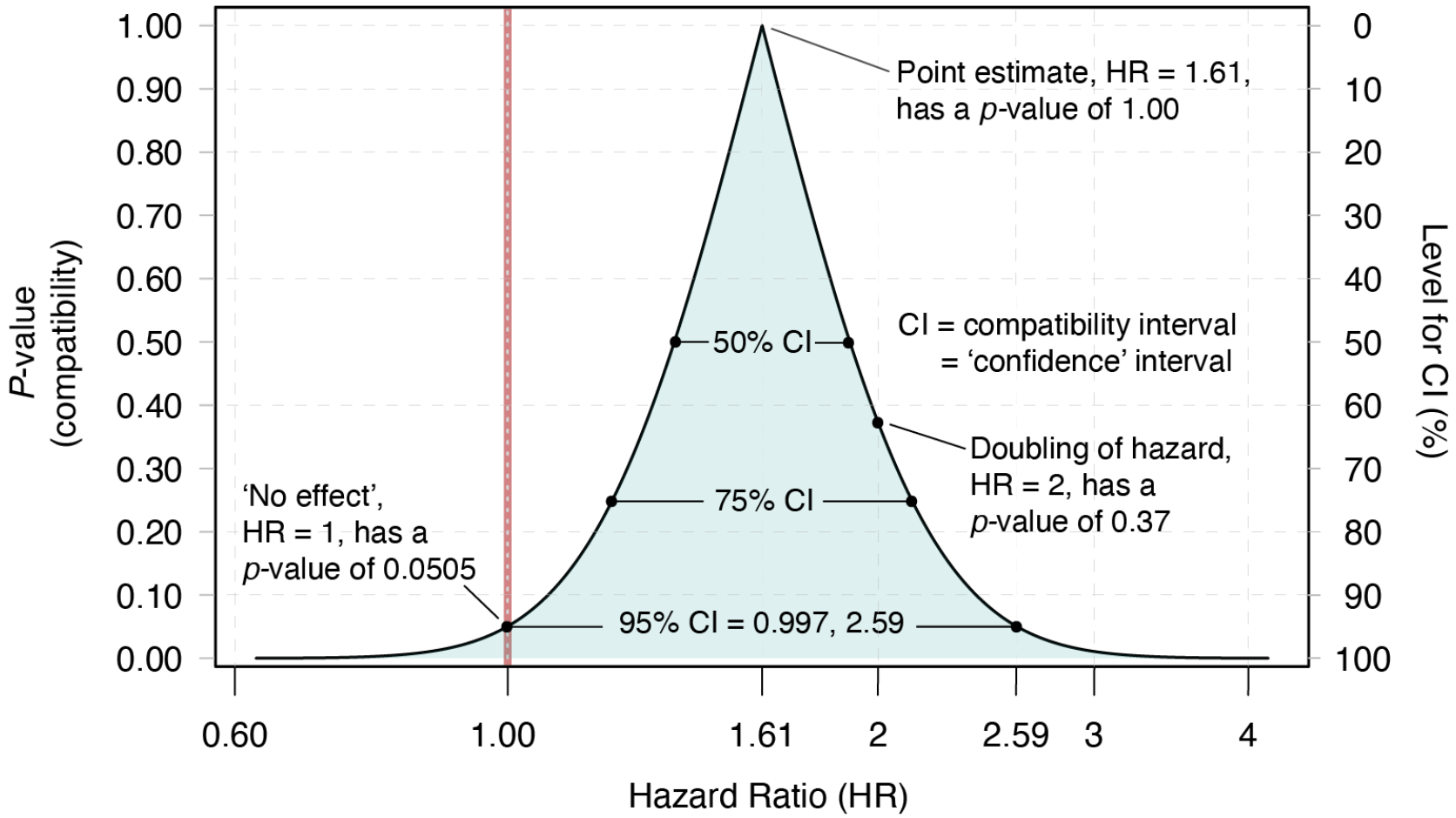
JAMA | Original Investigation 2017

# Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children

**RESULTS** There were 35 906 singleton births at a mean gestational age of 38.7 weeks (50.4% were male, mean maternal age was 26.7 years, and mean duration of follow-up was 4.95 years). In the 2837 pregnancies (7.9%) exposed to antidepressants, 2.0% (95% CI, 1.6%-2.6%) of children were diagnosed with autism spectrum disorder. The incidence of autism spectrum disorder was 4.51 per 1000 person-years among children exposed to antidepressants vs 2.03 per 1000 person-years among unexposed children (between-group difference, 2.48 [95% CI, 2.33-2.62] per 1000 person-years; hazard ratio [HR], 2.16 [95% CI, 1.64-2.86]; adjusted HR, 1.59 [95% CI, 1.17-2.17]). After inverse probability of treatment weighting based on the high-dimensional propensity score, the association was not significant (HR, 1.61 [95% CI, 0.997-2.59]). The association was also not significant when exposed children were compared with unexposed siblings (incidence of autism spectrum disorder was 3.40 per 1000 person-years vs 2.05 per 1000 person-years, respectively; adjusted HR, 1.60 [95% CI, 0.69-3.74]).

**CONCLUSIONS AND RELEVANCE** In children born to mothers receiving public drug coverage in Ontario, Canada, in utero serotonergic antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child. Although a causal relationship cannot be ruled out, the previously observed association may be explained by other factors.

# Compatibility graph (P-value function)




Amrhein & Greenland 2022; Rafi & Greenland 2020

A more correct description:

"Given our statistical model, our estimate was a 61% hazard increase.

However, under the same model, every hypothesis from no increase up to a 159% hazard increase was reasonably **compatible** with our data.

Thus, while quite imprecise, these results are most consistent with previous observations of a positive association."

Amrhein, Trafimow, Greenland 2019; Rafi & Greenland 2020

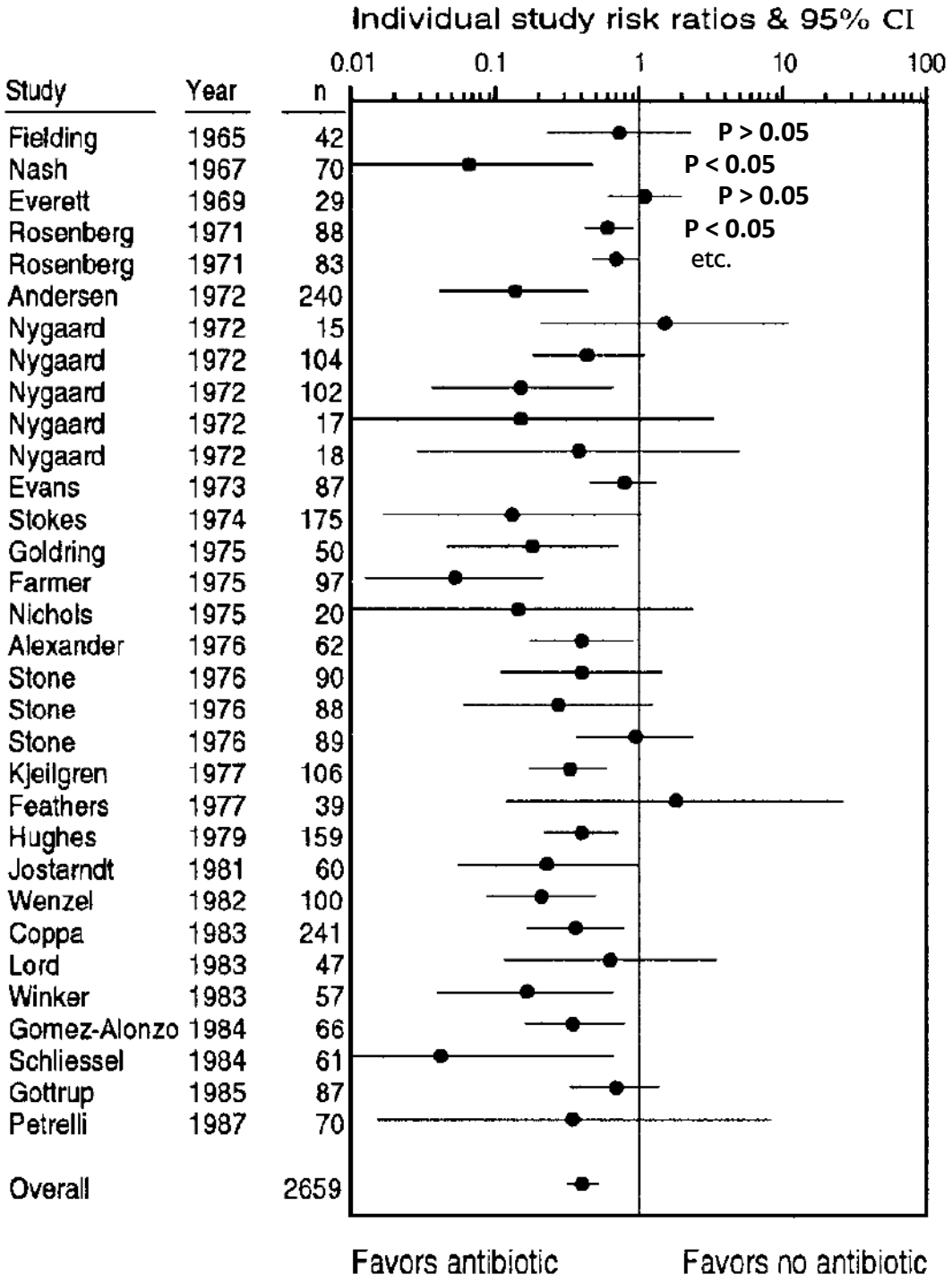
# Antidepressants in Pregnancy: No Link to Autism, ADHD

Batya Swift Yasgur, MA, LSW

April 21, 2017

Use of antidepressants before and during pregnancy does not cause autism, or attention-deficit/hyperactivity disorder (ADHD) new research shows.

"Medscape is the leading online global destination for physicians and healthcare professionals worldwide"



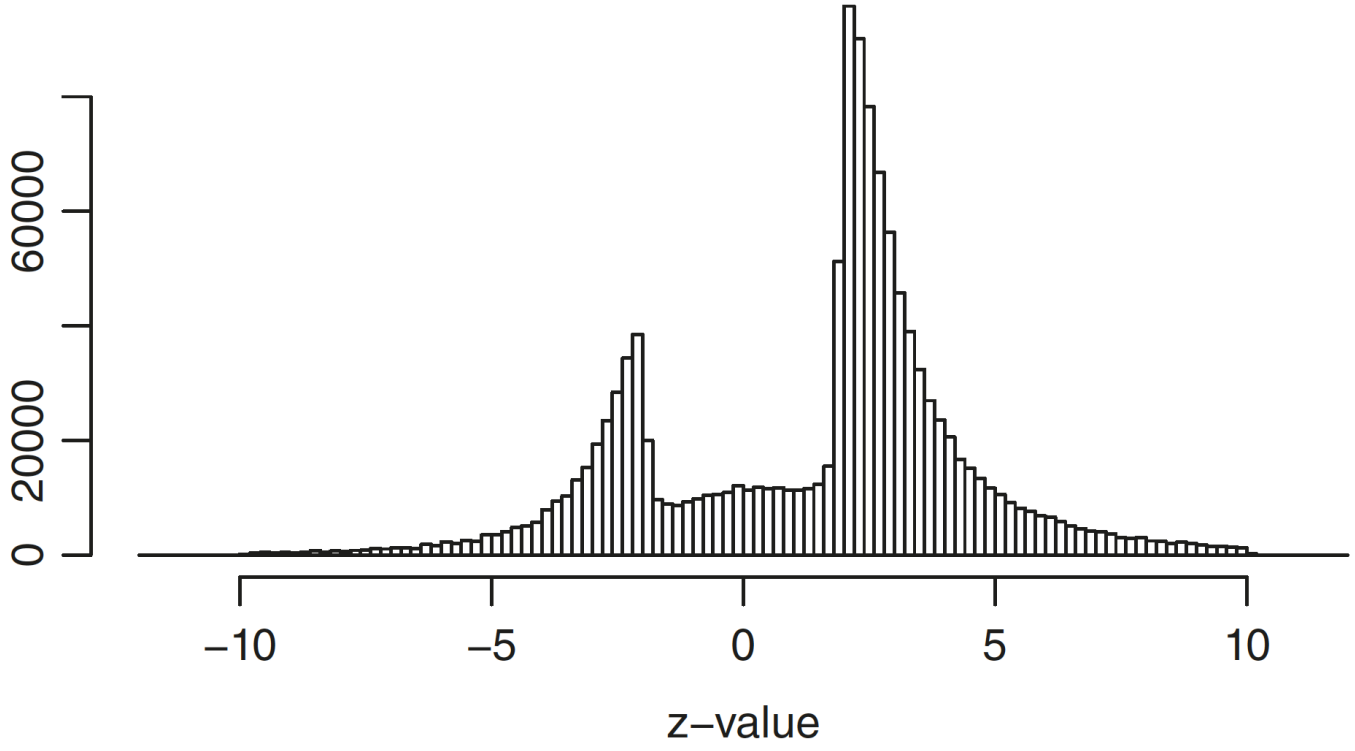

Results vary from study to study:
Single studies are not trustworthy, whether they are 'significant' or not.

Scientific generalization requires replication and meta-analysis including half of the studies that were not 'statistically significant'.

Studies investigating antibiotic prophylaxis compared with no treatment in colon surgery.
Analysed outcome: wound infection.
Ioannidis & Lau 1999

# Publication bias

'Non-significant' estimates (standard scores within ± 2 SD from the mean) are not published

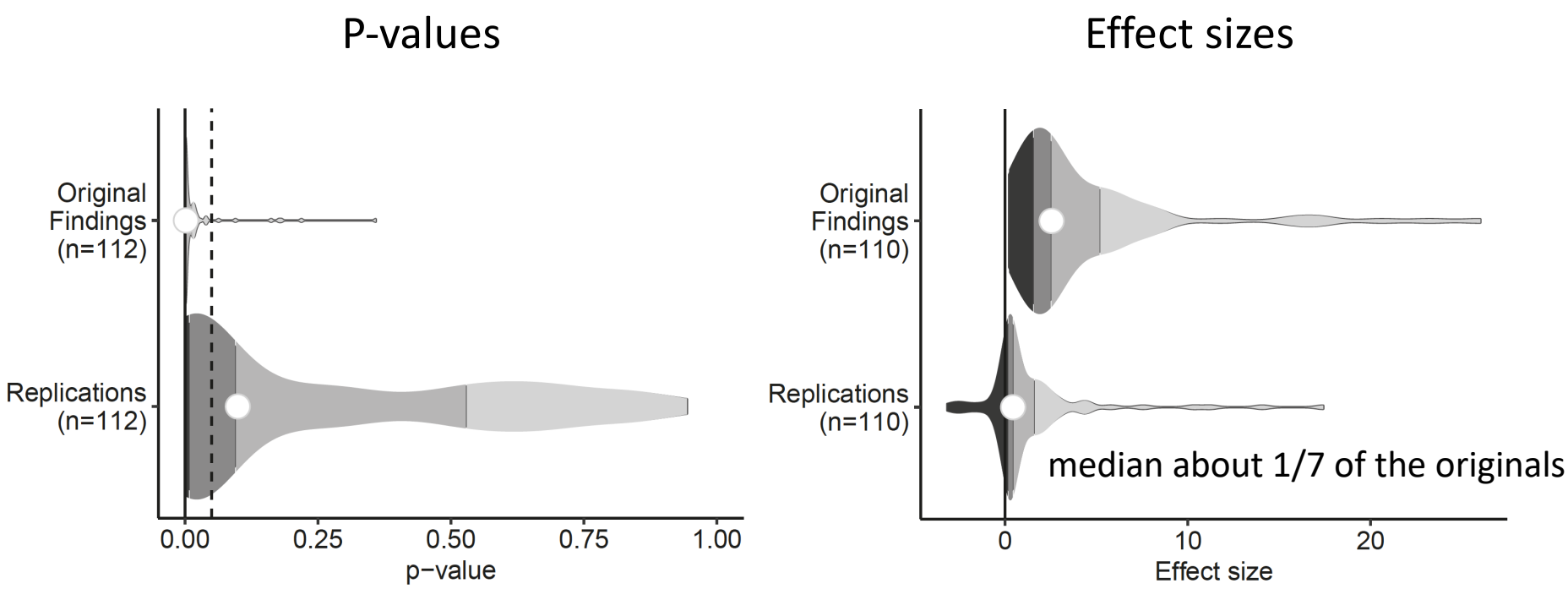



The distribution of 1.3 million results from Medline (1976–2019)
van Zwet & Cator 2021, Statistica Neerlandica

# Effect size inflation / Truth inflation / Winner's curse

Usually only the largest effects will become significant

⇨ significant effect sizes are almost always biased upwards




Replications of originally 'positive' results from preclinical cancer biology

Errington et al. 2021, eLife

Setting the agenda in research

# Comment

ILLUSTRATION BY DAVID PARKINS

## Reproducibility: expect less of the scientific paper

Olavo B. Amaral & Kleber Neves

Make science more reliable by placing the burden of replicability on the community, not on individual laboratories.

In 2018, we embarked on a journey to assess the reproducibility of biomedical research papers from Brazil. Thus began a multicentre collaboration of more than 60 laboratories to replicate 60 experiments from 2 decades of Brazilian publications[1]. We randomly selected experiments that used three common laboratory techniques: the MTT assay for cell viability, RT-PCR to measure specific messenger RNAs and the elevated plus maze to assess anxiety in rodents.

Each experiment will be repeated in three labs, and each lab has developed replication protocols based on the original article's written methods. The process of building, reviewing and preregistering these protocols has taken months of communication between the coordinating team and the labs performing replications. We had intense arguments around the meaning of positive and negative controls and the merits of different metrics to define replication success. We also spent many hours on mundane tasks, such as studying the nutritional content of different brands of bologna sausage to better emulate a cafeteria diet fed to rats in one experiment.

These are just some of the obstacles we





"Articles by individual research groups should thus be regarded as preliminary by default. If the expectation is that results of every publication hold true in other settings, models or populations, a reproducibility crisis seems inevitable."

Amaral & Neves 2021, Nature

"Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication"

Amrhein, Trafimow, Greenland 2019, The American Statistician


*"My question is: Are we making an impact?"*


# Let's embrace uncertainty
# Statistical significance is dead

48 papers from the 2020 volume of the Journal of Evolutionary Biology:

The results sections of the papers presented 49 significance tests on average (median 23, range 0–390).

No study presented a pre-specified (or pre-registered) alternative hypothesis, power calculation and the probability of 'false negatives' (beta error rate).

We conclude that studies in ecology and evolutionary biology are mostly exploratory and descriptive.

"Why and how we should join the shift from significance testing to estimation"
Berner & Amrhein 2022, Journal of Evolutionary Biology

| | Country | Popul | Family | Sex | Temp | Country*Temp |
|---|---|---|---|---|---|---|
| Larval time | | | *** | *** | *** | *** |
| Pupal time | | | *** | | *** | |
| Pupal mass | | *** | *** | *** | *** | *** |
| Larval growth rate | | *** | | *** | *** | *** |
| Adult mass | | | *** | *** | *** | |
| Thorax mass | | | *** | *** | *** | |
| Abdomen mass | | | *** | *** | *** | *** |
| Thorax/Abdomen | | | | *** | *** | *** |
| Forewing area | | *** | *** | *** | *** | |
| FW melanization | *** | | *** | *** | *** | *** |
| Hindwing area | | *** | *** | *** | *** | |
| HW melanization | *** | | *** | *** | *** | |
| FW-HW ratio | | *** | *** | *** | | |
| Wing loading | | | *** | *** | *** | |
| Wing aspect ratio | | | | *** | | |
| Heat tolerance | | | *** | *** | *** | |
| PMA | | | | | *** | |

102 significance tests, no P-values, no effect estimates

The replication crisis in science is **not** the product of the publication of unreliable findings.

The publication of unreliable findings is unavoidable: as the saying goes, if we knew what we were doing, it would not be called research.

Rather, the replication crisis has arisen because unreliable findings are presented as reliable.

Amrhein, Gelman, Greenland, McShane 2019, PeerJ Preprints

Swiss Reproducibility Network

The SwissRN is a peer-led consortium that aims to promote and ensure rigorous research practices in Switzerland

# Swiss Reproducibility Conference 2024

Monday 10.6. – Tuesday 11.6.2024, Zürich

www.reproducibility.ch

https://camargue.unibas.ch/en/reproducibility