NYU HEALTH SCIENCES LIBRARY

Replicability 2

Fred LaPolla, MLS
Alisa Surkis, PhD MLS

NYU HEALTH SCIENCES LIBRARY

Review!

List factors that limit replicability

NYU HEALTH SCIENCES LIBRARY

0

Why does low power limit replicability?

NYU HEALTH SCIENCES LIBRARY

Why might a lack of blinding and randomization limit replicability?

NYU HEALTH SCIENCES LIBRARY

0

NYU HEALTH SCIENCES LIBRARY

Agenda

- Sex as a Biological Variable
- Misunderstood Statistics
- Publication Bias & Transparent Reporting
- The NIH

NYU HEALTH SCIENCES LIBRARY

Objectives: Students will be able to...

- → Explain why it is important to account for sex differences
- → Define p means and discuss problems with "significance"
- → Name ways that transparency can be improved in scientific reporting

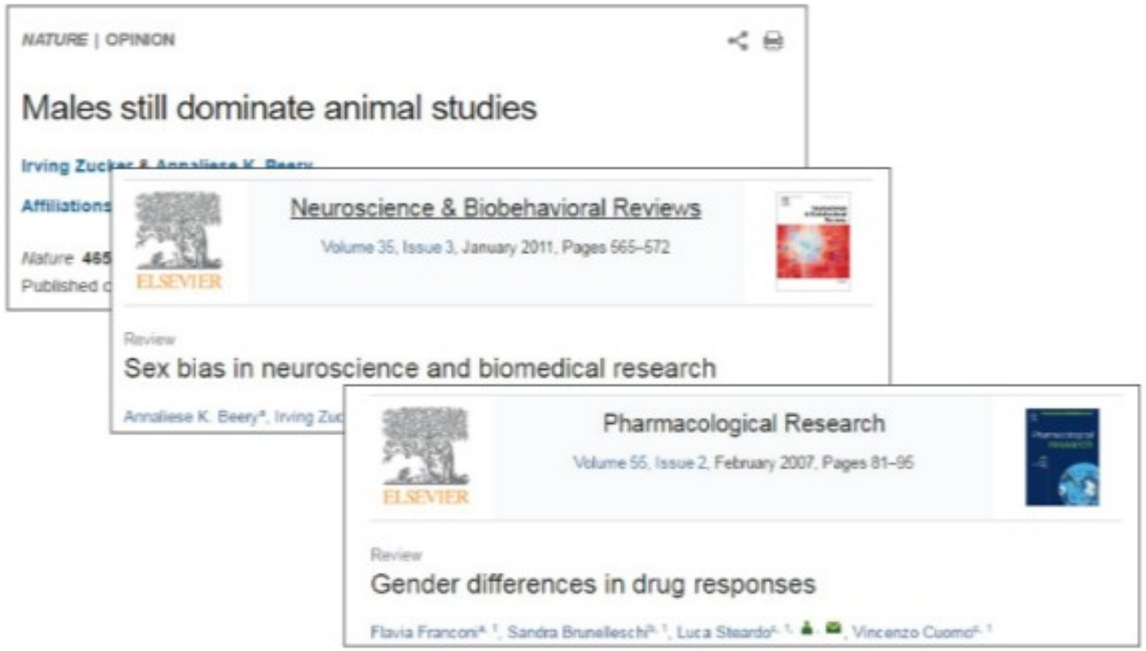
NYU HEALTH SCIENCES LIBRARY

Sex as a Biological Variable

What aspects of scientific research may be impacted by sex as a variable (what may be different and why could it matter to research)?

NYU HEALTH SCIENCES LIBRARY

0

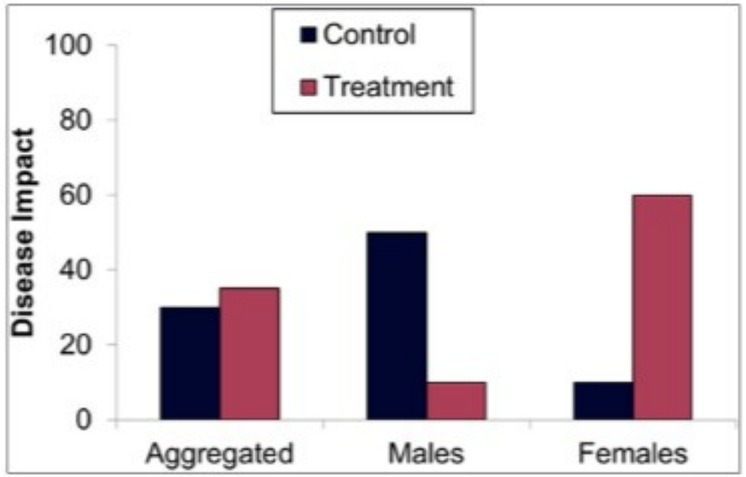


Sex as a biological variable matters due to differences in:

- → Dosing
- → Adverse Events
- → Intervention/Stimulus Response
- → Disease Presentation

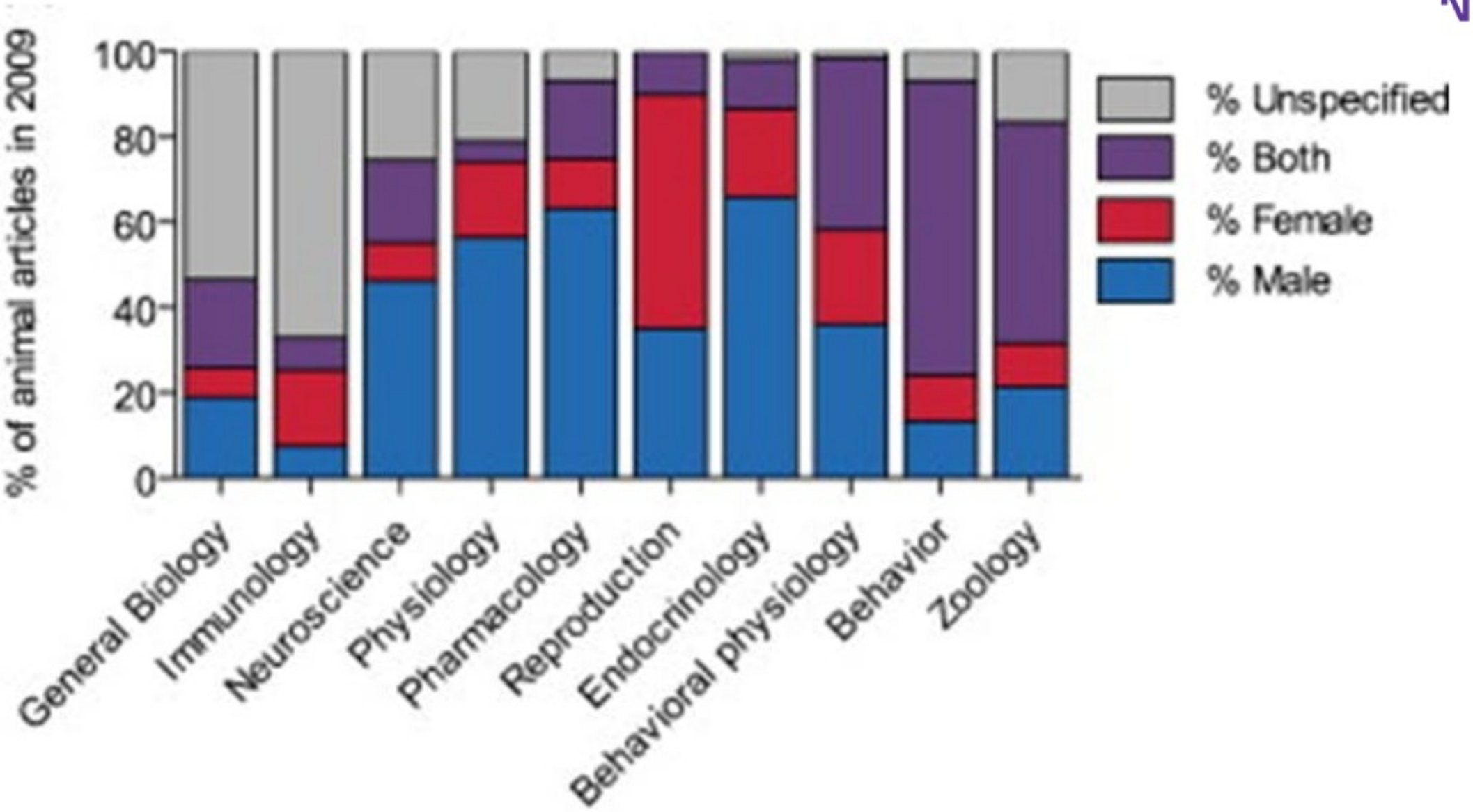
Implications

- 10 drugs removed from the U.S. market between 1997-2000
- Eight withdrawn because of side effects occurring only or primarily in women

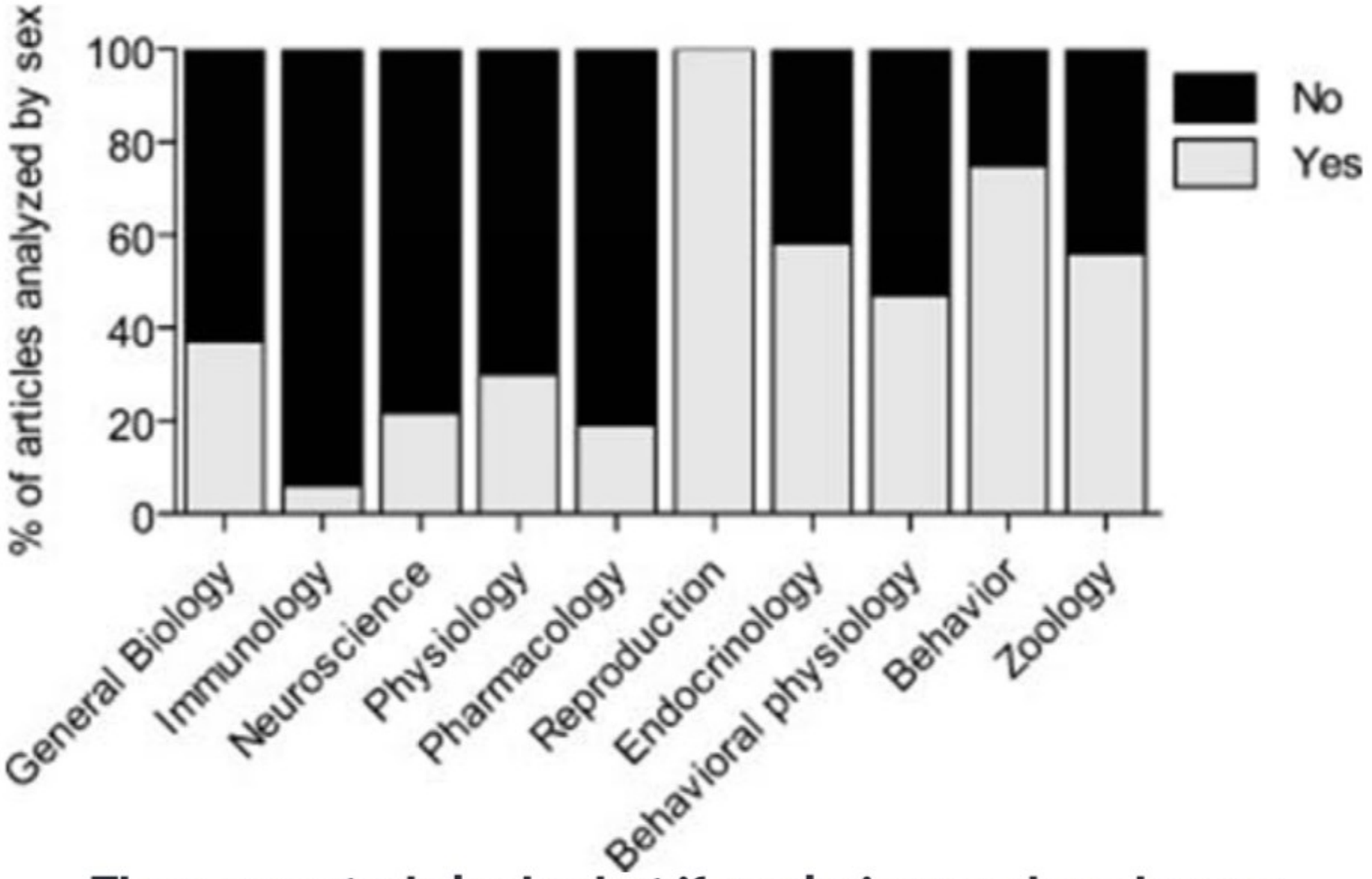


This relates to replicability because

- We must know the range of validity of results:
- Drug X has some effect Y
- Drug X has some effect Y in species Z
- Drug X has some effect Y in males of species Z



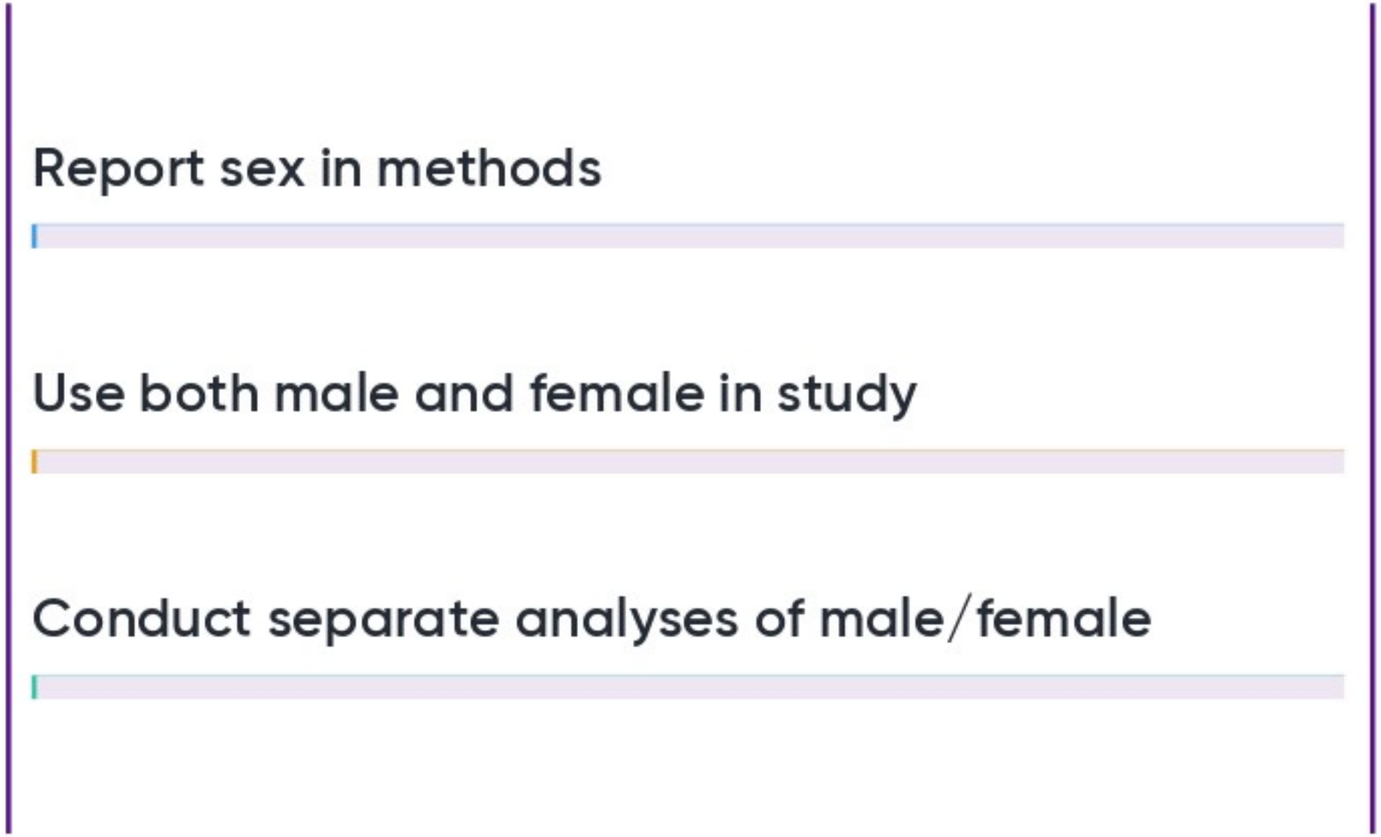
A 2011 Study examined distribution of sex in 10 biological fields



The same study looked at if analysis was done by sex

Estimate the percentage of time each was done in studies you've been a part of

NYU HEALTH SCIENCES LIBRARY



0

Ways to Consider Sex as a Biological Variable

- Add terms like "sex", "male" or "female" to your literature search to consider known sex differences
- If the literature search causes you to suspect sex differences, conduct pilot studies to determine whether powering by sex differences is warranted
- Randomize and balance the sexes
- Disaggregate data to see if there are sex differences

NYU HEALTH SCIENCES LIBRARY

Misunderstanding and Misusing Statistics

What is a p value?

NYU HEALTH SCIENCES LIBRARY

P value

The probability of obtaining a result as large or extreme as the results of a study if the null hypothesis were true.

|

If p = 0.05, we are saying that if we ran the study 100 times, we would expect 5 times out of the hundred we would get a result as "far" from the null hypothesis' value if the null hypothesis were true

?

P-hacking

- Running many statistical tests until significance is achieved
- Transforming the data until significance is achieved
- Non-transparent inclusion criteria
- Running tests on part of the data and stopping once significance is found
- Cherry Picking: Only reporting significant findings

Exercise

We are going to run a test called a t-test on 10 random numbers. A t-test is a simple test for comparing two means of two groups to see if there is a difference.

Why is it not enough to simply compare the means of two groups and tell if the difference is meaningful?

NYU HEALTH SCIENCES LIBRARY

Exercise Jargon

A standard deviation is a measure of how much variance exists in a set of data. It tells us how "far" values in our sample are from the mean.

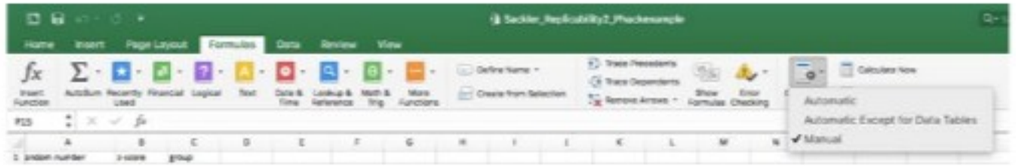
|

A Z-Score or standard score is a normalization of an individual measurement in terms of standard deviations. The z-score tells you how many standard deviations away from the mean (0) a given value is.

|

A z score of +/- 1.96 corresponds to 95% of values falling within that range.

?



First: Open Excel

- Go to the Formulas tab
- Open the menu labeled Calculation Options
- Change it to "Manual"
- Now to run a calculation, we will manually check "Calculate Now"

Second: Create 10 random numbers

- In cell A1, write "Random Number"
- In cells A2-A11, write"=rand()"
- This will generate a random number.

Third: Normalize your ten random numbers

- In cell b1, type z-score
- In b2, type "=norm.s.inv(a2)"
- Copy this down to b11, so that each cell has a z score for its neighboring random number
- Press "Calculate Now" in the formulas tab

Fourth: Put them in groups

- → In C1, write "Group"
- → Call C2:C6 1 or A
- → Call C7:C11 2 or B

Fifth (and final): Run a t-test to compare the difference between the two groups

- In an empty cell, type: "=t.test(b2:b6,b7:b11, 2,2)"
- This means, run a t-test comparing the group of b2-b6 to b7-b11, test for two tails and assume equal variance. Tails refer to "direction" of effect
- Run the test 20 times by hitting **Calculate Now**
- **Count** how many times the results are <0.05

How many times did you get a value < .05?

NYU HEALTH SCIENCES LIBRARY

Try it with an actual difference

- In slide b7, change the equation to =norm.s.inv(a7)+0.5
- Copy this down through B11
- Run the test 20 times by hitting **Calculate Now**
- **Count** how many times the results are <0.05

How many times did you get a value < .05?

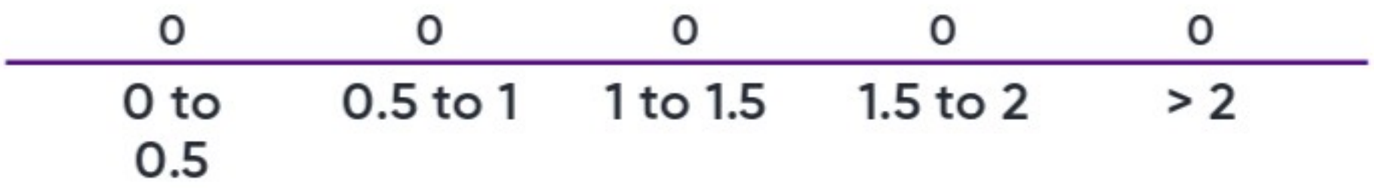
NYU HEALTH SCIENCES LIBRARY

Winner's Curse

- In any cell type = ABS(AVERAGE(B7:B11) – AVERAGE(B2:B6))
- Run Calculate Now until you get a value of $p < 0.05$
- Keep track of that value of the difference of means (effect size)

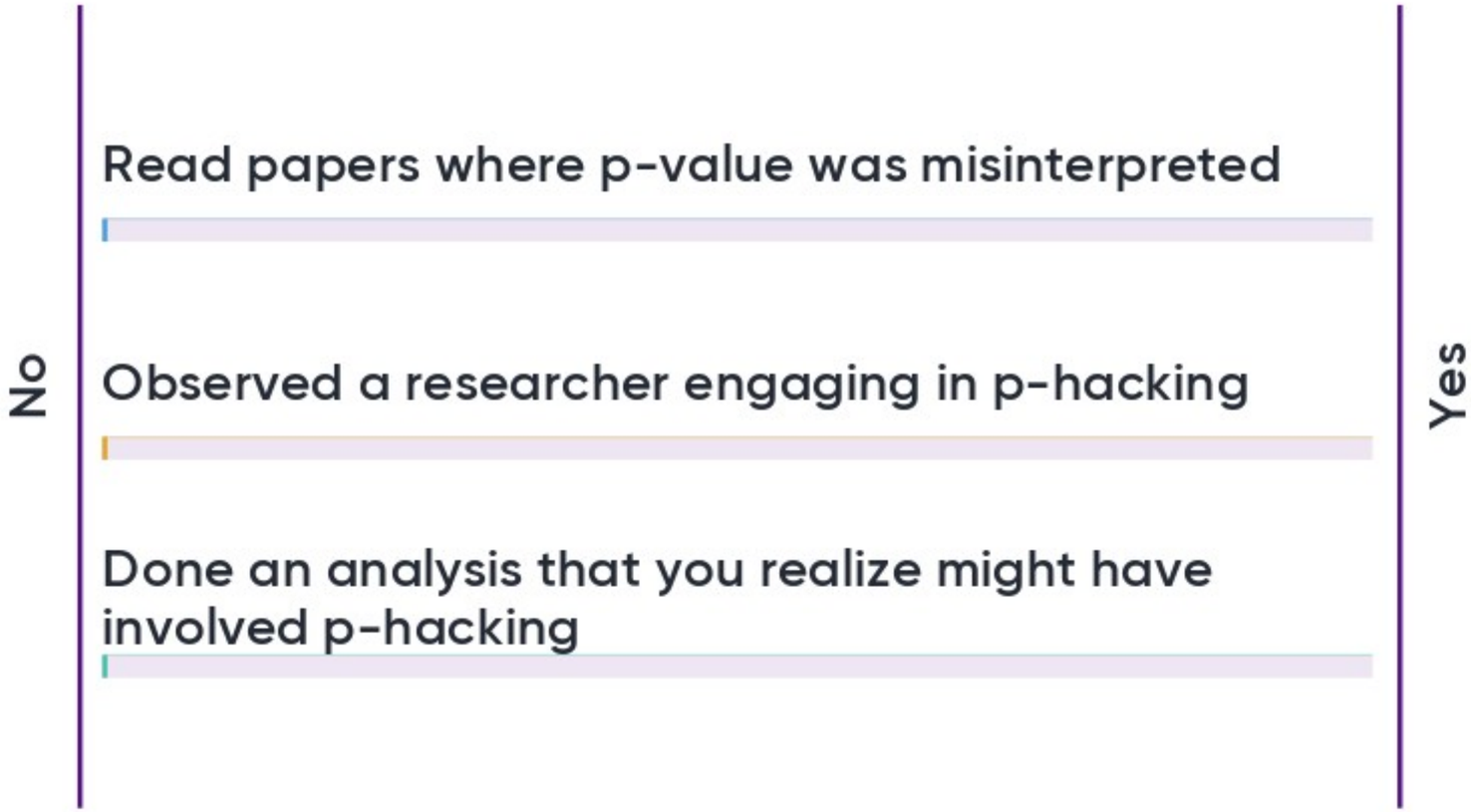
How big was the difference in means when you got p < .05?

NYU HEALTH SCIENCES LIBRARY



0

Have you done/observed any of the following?

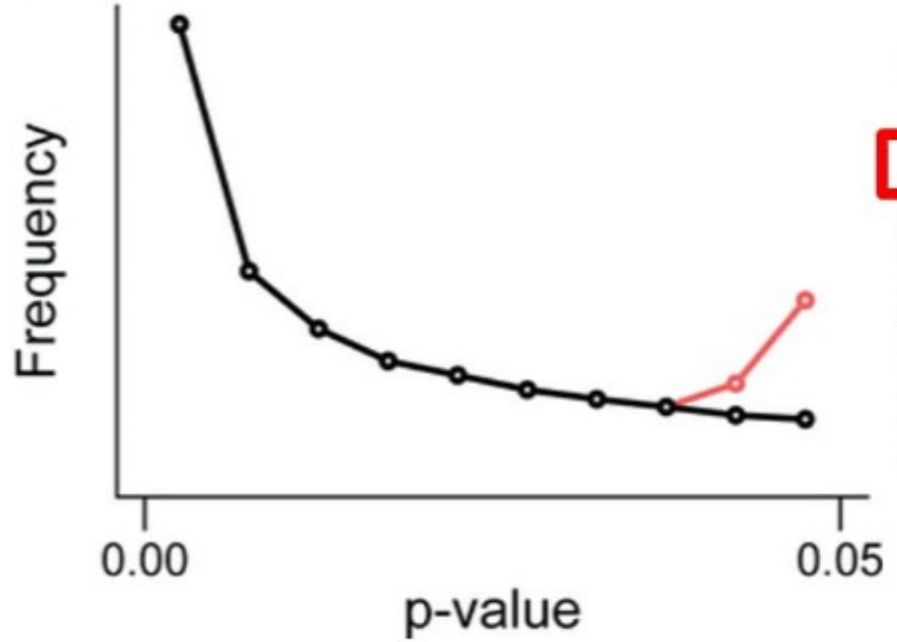


PLOS | BIOLOGY

The Extent and Consequences of P-Hacking in Science

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions

Published: March 13, 2015 • https://doi.org/10.1371/journal.pbio.1002106



Discipline	Number of *p*-values between 0 and 0.025	Number of *p*-values between 0.025 and 0.05	Binomial test for evidential value	Number of *p*-values between 0.04 and 0.045	Number of *p*-values between 0.045 and 0.05	Binomial test forp-hacking
Medical and health sciences	45,460	16,537	<0.001	1,477	1,785	<0.001
Multidisciplinary	21,209	6,793	<0.001	638	750	0.001
Psychology and cognitive sciences	1,355	487	<0.001	29	50	0.012
Studies in human society	139	45	<0.001	8	3	0.967
Technology	94	37	<0.001	3	3	0.656

Number of *p*-values in each bin is the mean number based on 1,000 bootstraps of one *p*-value per Results section, rounded to the nearest whole number. Disciplines (n = 8) for which we found fewer than 50 *p*-values below 0.05 in the Results section were excluded.

doi:10.1371/journal.pbio.1002106.t001

2015 study found overabundance of p values just under 0.05, implying p-hacking

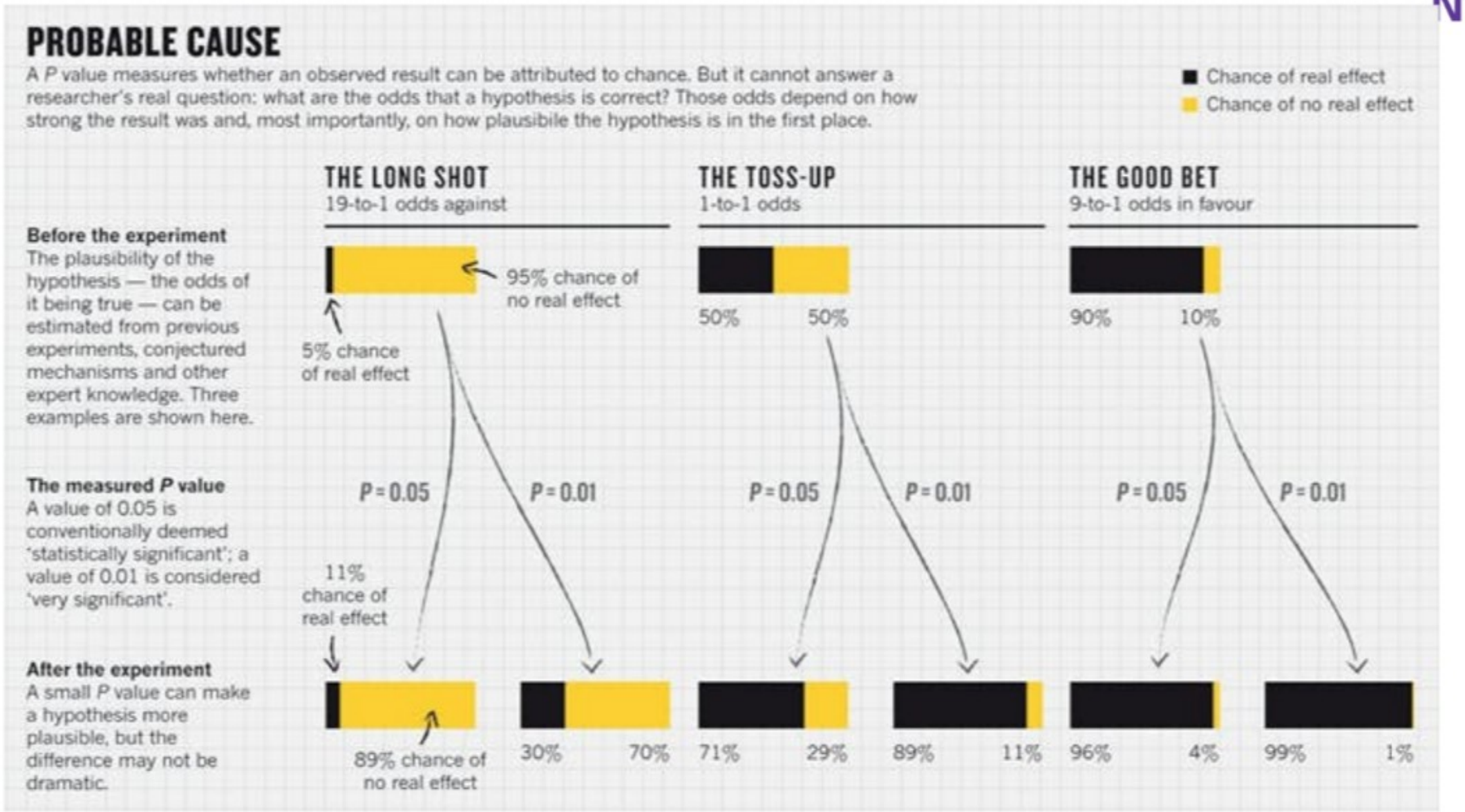
NYU HEALTH SCIENCES LIBRARY

P values

Even without p-hacking, misinterpretation leads to problems

P Values

- Introduced by Ronald Fisher in the 1920s as a general guideline
- Not meant as a binary
- The more implausible a hypothesis the greater the chance that a finding is a false positive, regardless of p value



The odds of a hypothesis being true are different than the p value

nature human behaviour

Comment

Justify your alpha

Daniel Lakens ✉, Federico G. Adolfi, [...] Rolf A. Zwaan

In response to recommendations to redefine statistical significance to $P \leq 0.005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level.

https://www.nature.com/articles/s41562-018-0311-x

Some researchers have called for a smaller alpha, others disagree

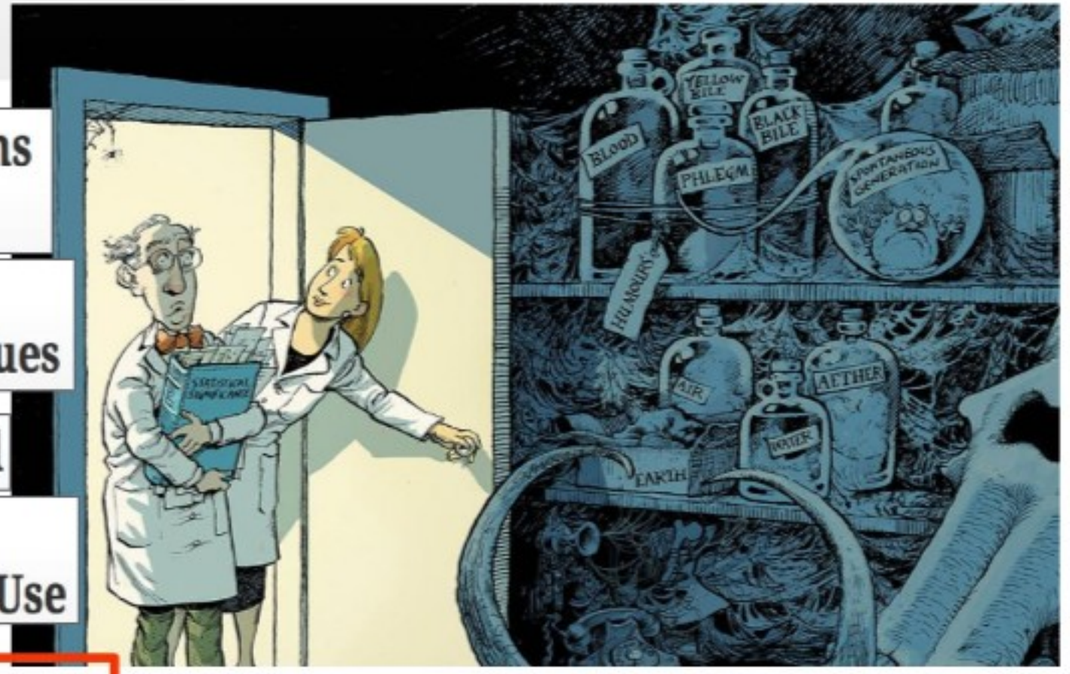
Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values

The False Positive Risk: A Proposal Concerning What to Do About *p*-Values

The *p*-Value Requires Context, Not a Threshold

A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting its Use

Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility



COMMENT · 20 MARCH 2019

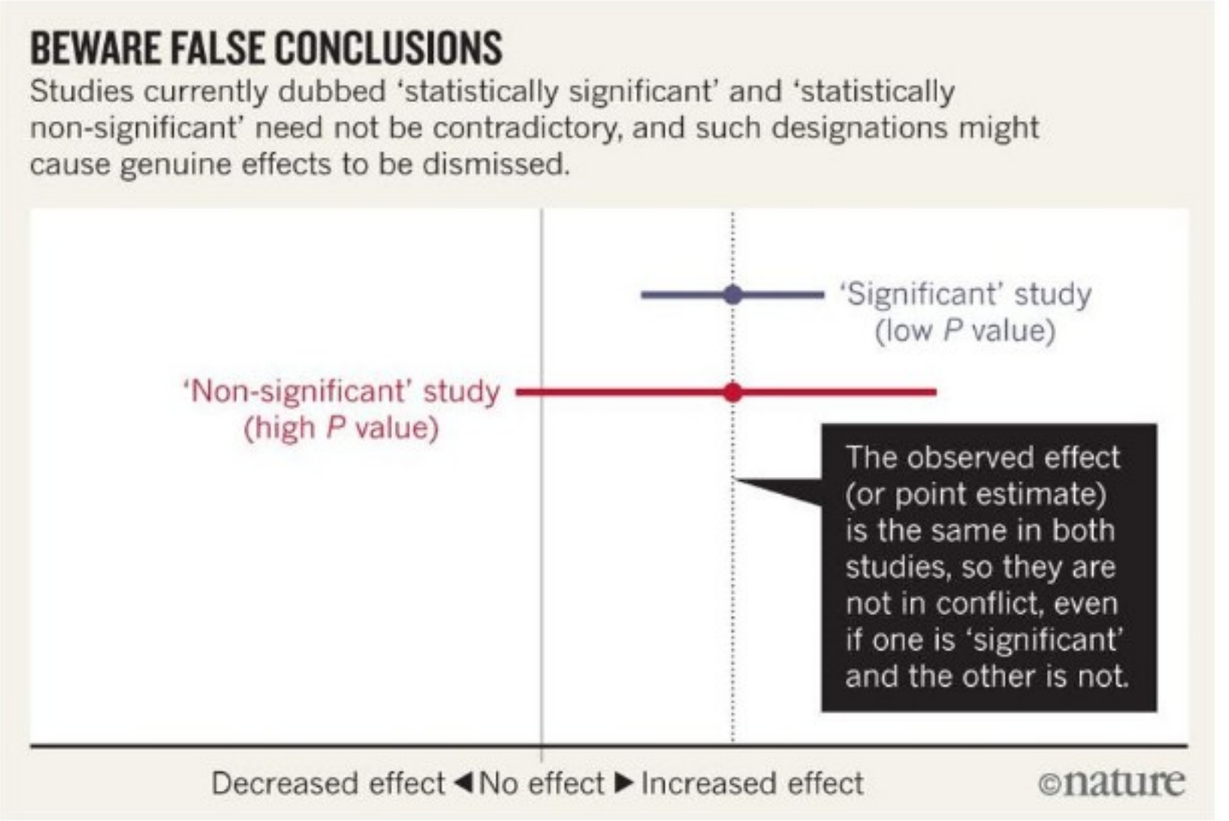
Scientists rise up against statistical significance

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.

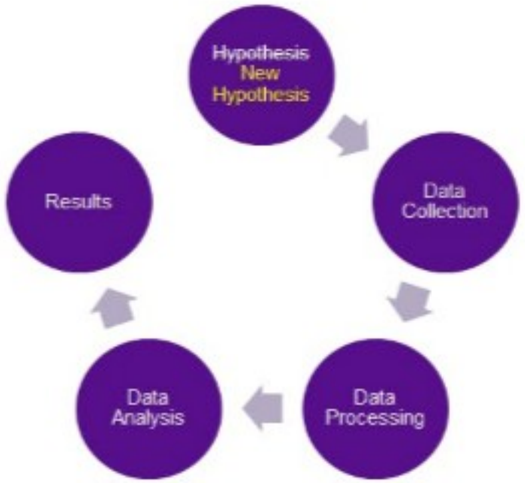
Growing chorus of voices calling for moving away from "significance"

Scientists rise up against statistical significance

- From study of effects of unintended effects of anti-inflammatory drugs
- Study 1: **significant**
  Study 2: **non-significant**
- Conclusion that second study contradicted first



A review of 791 articles, found that 51% incorrectly described P Values

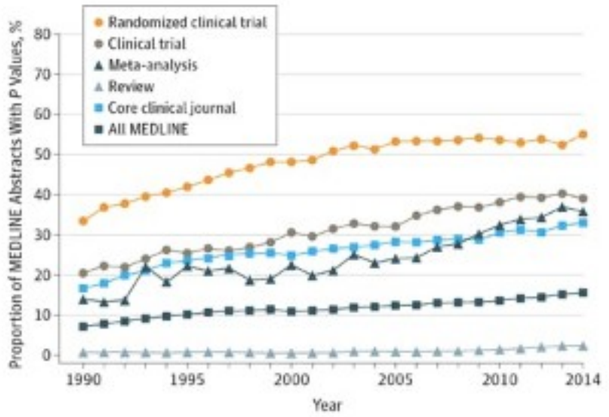


HARKing: Hypothesizing After Results are Known

- Normally a hypothesis is based on prior research, observations and exploratory analysis
- Data is collected to confirm and test the hypothesis
- HARKing refers to making a new hypothesis based on collected data in confirmatory research

NYU HEALTH SCIENCES LIBRARY

Publication Bias

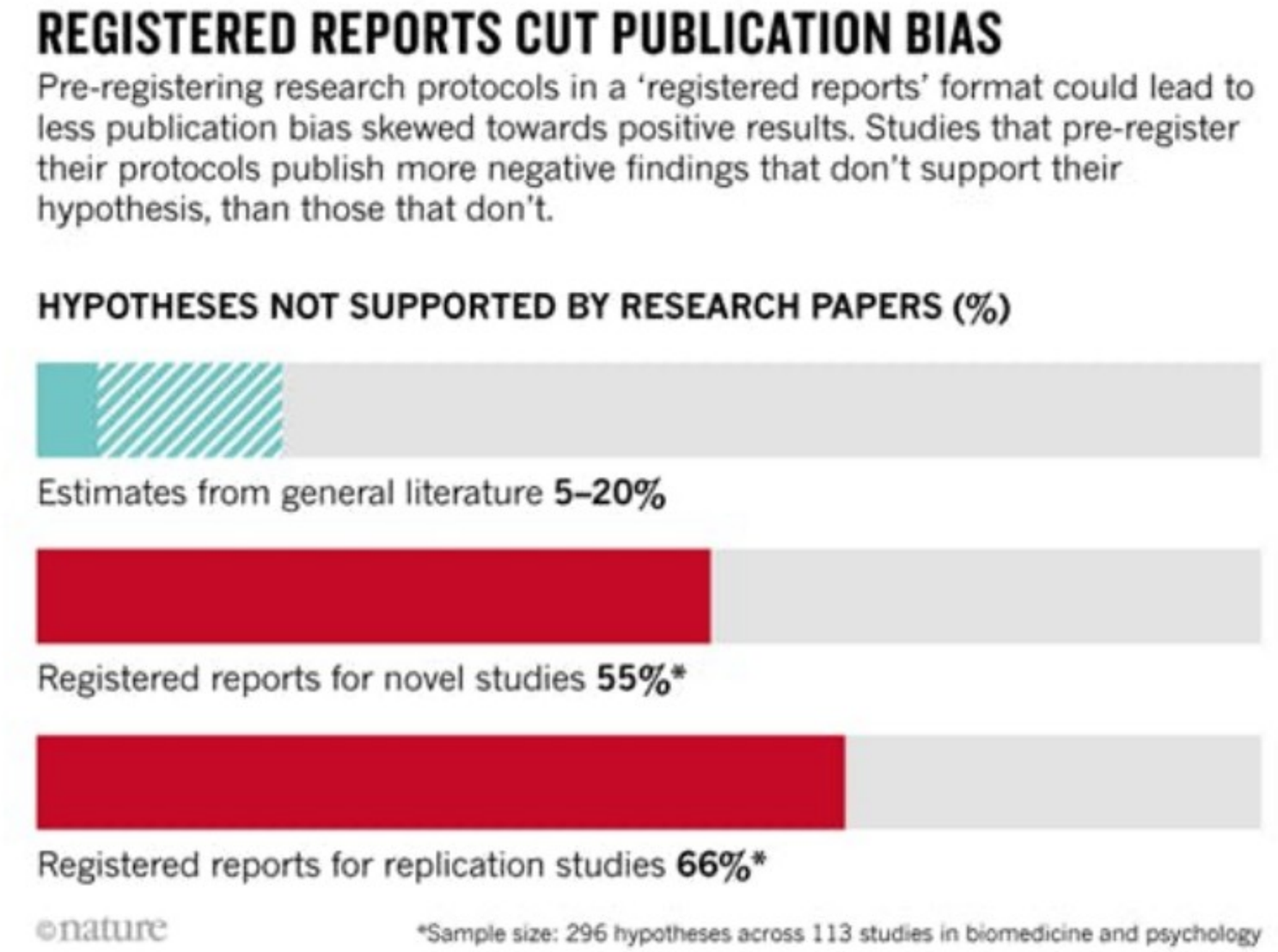


Positive findings more likely to be published

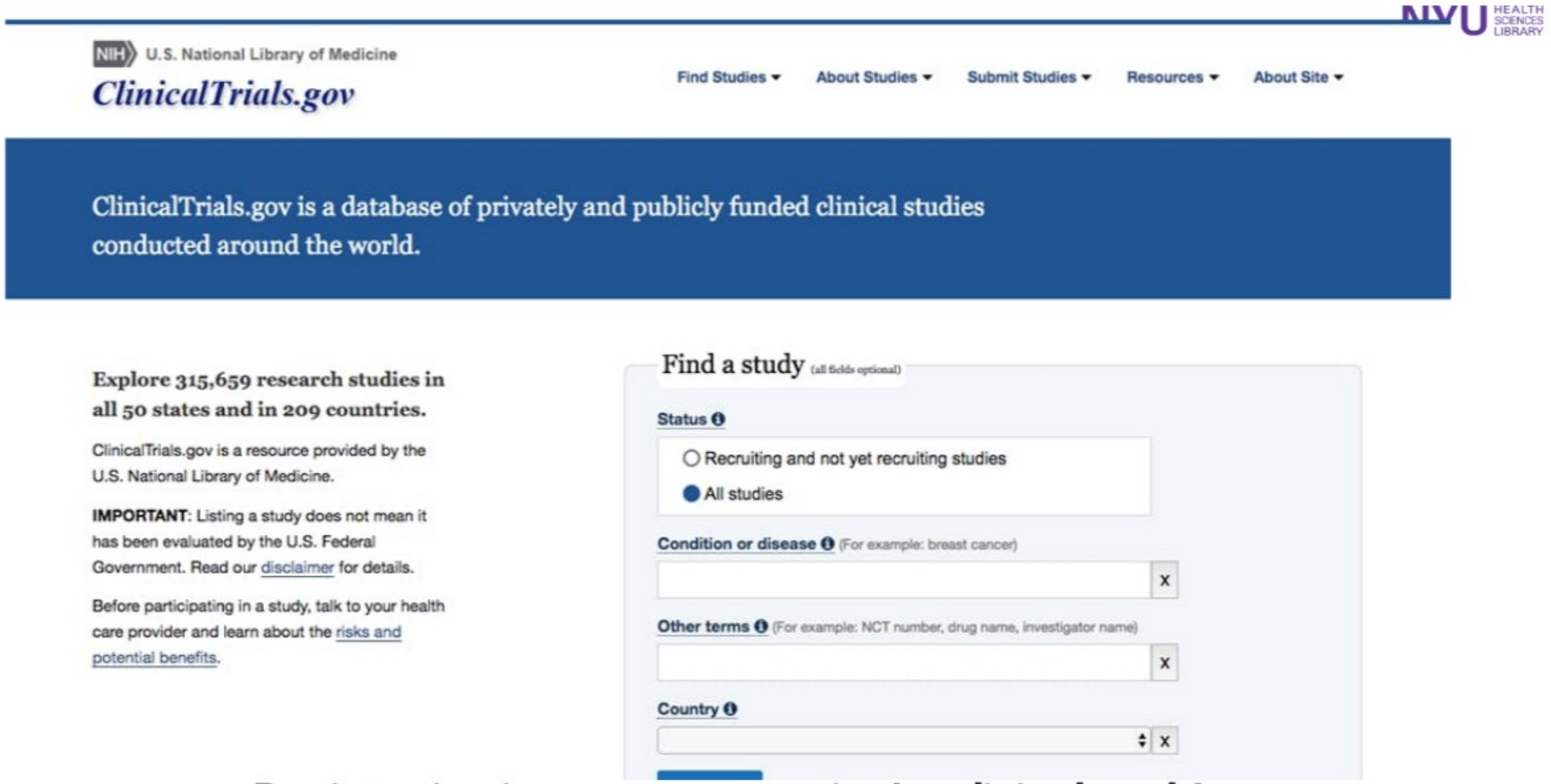
- A 2016 JAMA article analyzed papers 1990-2015
- Found 96% had at least one p value under 0.05

Possible solution: Registration

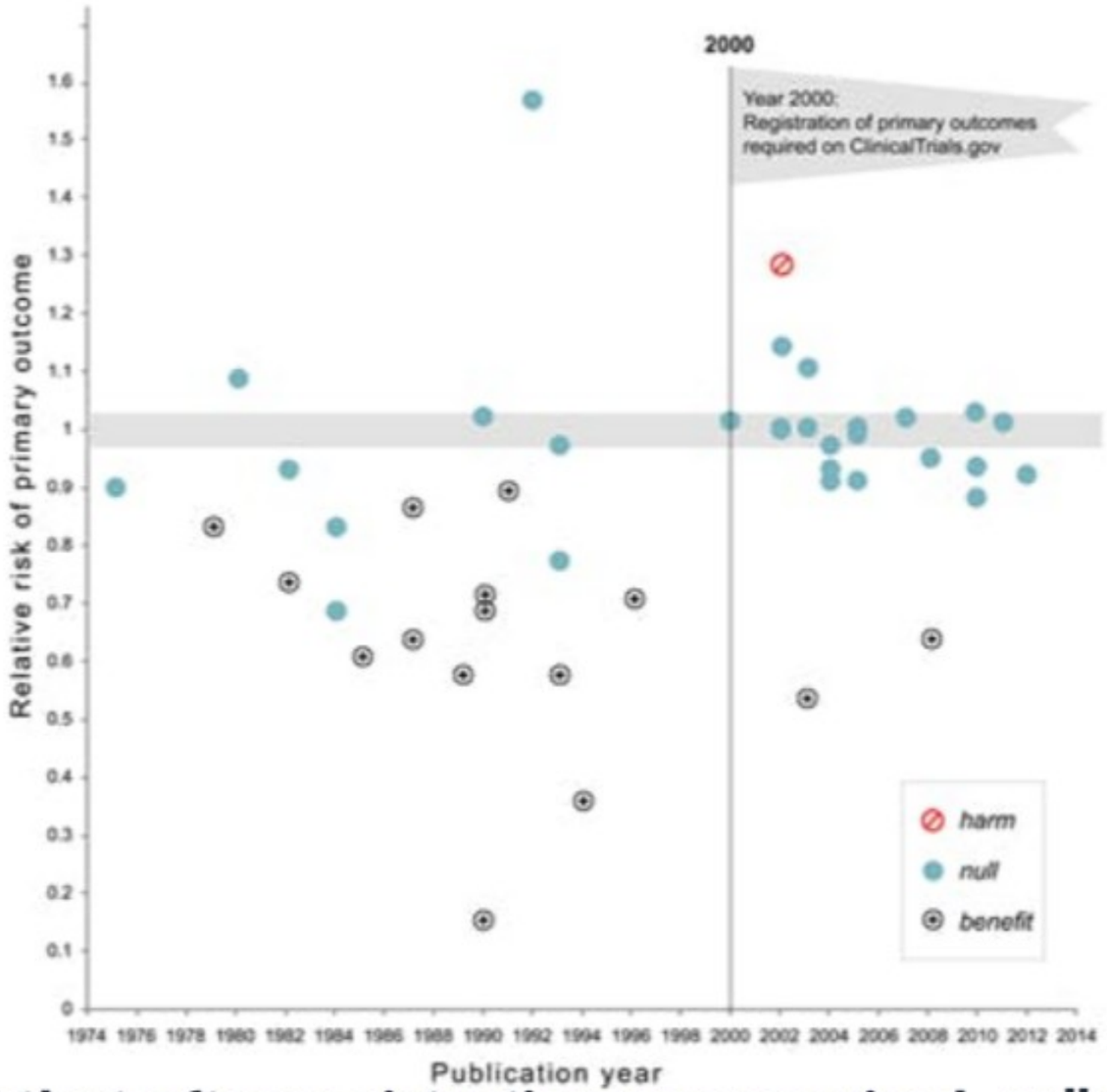
- Study proposal is reviewed before research
- Pre-registered proposals may be provisionally accepted before outcomes known
- Currently 207 journals have adopted registered reports



Registered reports have much higher rate of null findings



Registration is more common in the clinical world



PlosOne study found that after registration was required, null results became more common

Animal Registries

- Registration can boost credibility
- But remains relatively rare due to administrative burden

NYU HEALTH SCIENCES LIBRARY

Transparent Reporting

2009 study on 271 publications

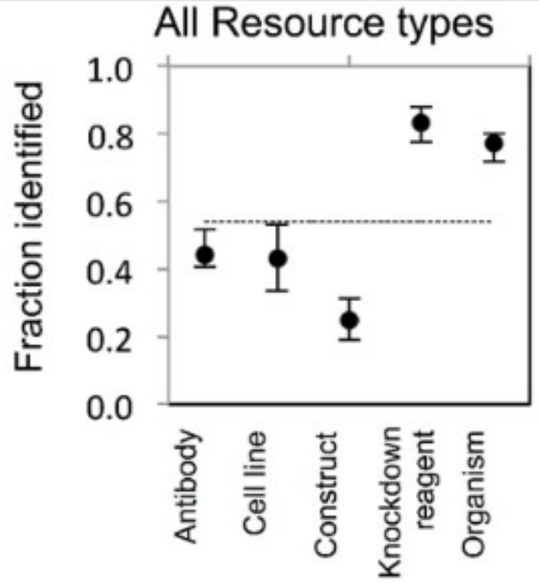
- → 41% did not report clear hypothesis or animal characteristics
- → 30% were unclear about statistical methods

2016 study on 47 preclinical studies on acute lung injury and mesenchymal stem cells

- 2% of studies provided sample size estimation info
- 4% provided inclusion/exclusion criteria
- 4% provided housing info
- 45% provided age range of animals
- 79% provided animal sex
- 74% reported stem cell dose

PeerJ

On the reproducibility of science: unique identification of research resources in the biomedical literature



2013 paper

- Analyzed 238 manuscripts from 84 journals (including Science, Nature)
- Summarized average fraction of resources uniquely identified

PERSPECTIVE

doi:10.1038/nature11556

A call for transparent reporting to optimize the predictive value of preclinical research

Story C. Landis[1], Susan G. Amara[2], Khusru Asadullah[3], Chris P. Austin[4], Robi Blumenstein[5], Eileen W. Bradley[6], Ronald G. Crystal[7], Robert B. Darnell[8], Robert J. Ferrante[9], Howard Fillit[10], Robert Finkelstein[11], Marc Fisher[12], Howard E. Gendelman[13], Robert M. Golub[14], John L. Goudreau[15], Robert A. Gross[16], Amelie K. Gubitz[3], Sharon E. Hesterlee[18], David W. Howells[19], John Hugenard[20], Katrina Kelner[21], Walter Koroshetz[1], Dimitri Krainc[22], Stanley E. Lazic[23], Michael S. Levine[24], Malcolm R. Macleod[25], John M. McCall[26], Richard T. Moxley III[27], Kalyani Narasimhan[28], Linda J. Noble[29], Steve Perrin[30], John D. Porter[3], Oswald Steward[31], Ellis Unger[32], Ursula Utz[1] & Shai D. Silberberg[1]

2012 Report calls for as a minimum reporting:

- → Sample size estimation
- → Whether/how animals were randomized
- → Whether investigators were blind to treatment
- → How data was handled

Starting 2013 Nature began implementing a reproducibility checklist

- Included items like sample size, randomization and blinding
- A 2018 survey found 49% felt checklists improved quality of research
- Those who felt it improved research stated it improved statistics reporting

Do you feel that checklists improve the quality of research? Why or why not?

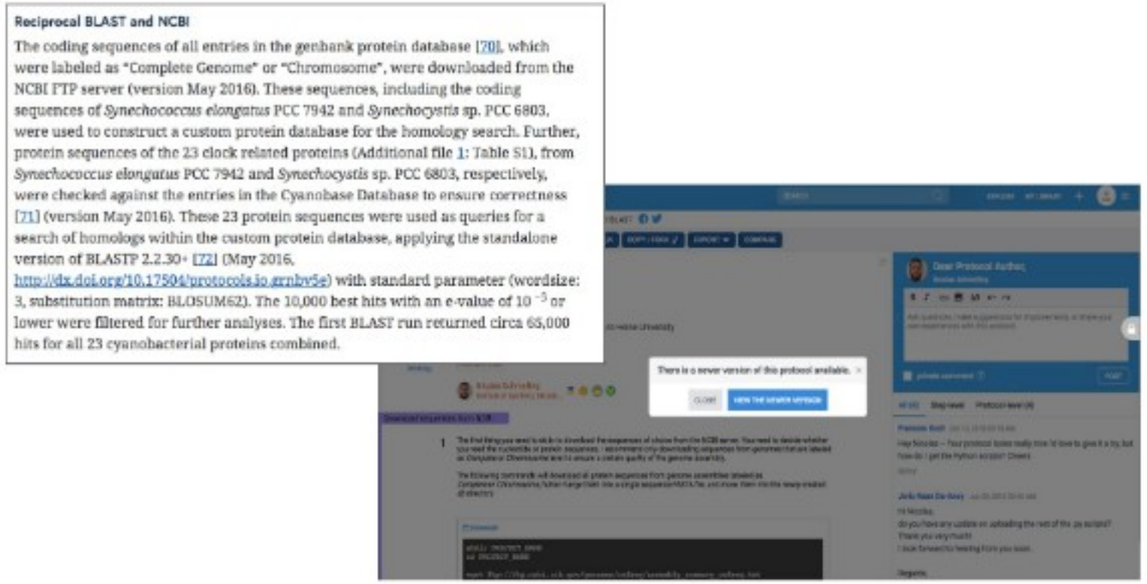
NYU HEALTH SCIENCES LIBRARY

0

Have you seen: We used a modified version of the protocol from Paper X

NYU HEALTH SCIENCES LIBRARY

0



Reciprocal BLAST and NCBI

The coding sequences of all entries in the genbank protein database [70], which were labeled as "Complete Genome" or "Chromosome", were downloaded from the NCBI FTP server (version May 2016). These sequences, including the coding sequences of *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803, were used to construct a custom protein database for the homology search. Further, protein sequences of the 23 clock related proteins (Additional file 1: Table S1), from *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803, respectively, were checked against the entries in the Cyanobase Database to ensure correctness [71] (version May 2016). These 23 protein sequences were used as queries for a search of homologs within the custom protein database, applying the standalone version of BLASTP 2.2.30+ [72] (May 2016, http://dx.doi.org/10.17504/protocols.io.grnbv5e) with standard parameter (wordsize: 3, substitution matrix: BLOSUM62). The 10,000 best hits with an e-value of 10^{-5} or lower were filtered for further analyses. The first BLAST run returned circa 65,000 hits for all 23 cyanobacterial proteins combined.

Following 1.5 years of a replicability attempt, Lenny Teytleman launched protocols.io

- → Open access repository of methods
- → Methods can be updated and linked
- → 10k protocols uploaded and 200 journals call on authors to link to protocols.io

Explore protocols.io at your table: have you seen it used? Have you used it, and would you? Why or why not?

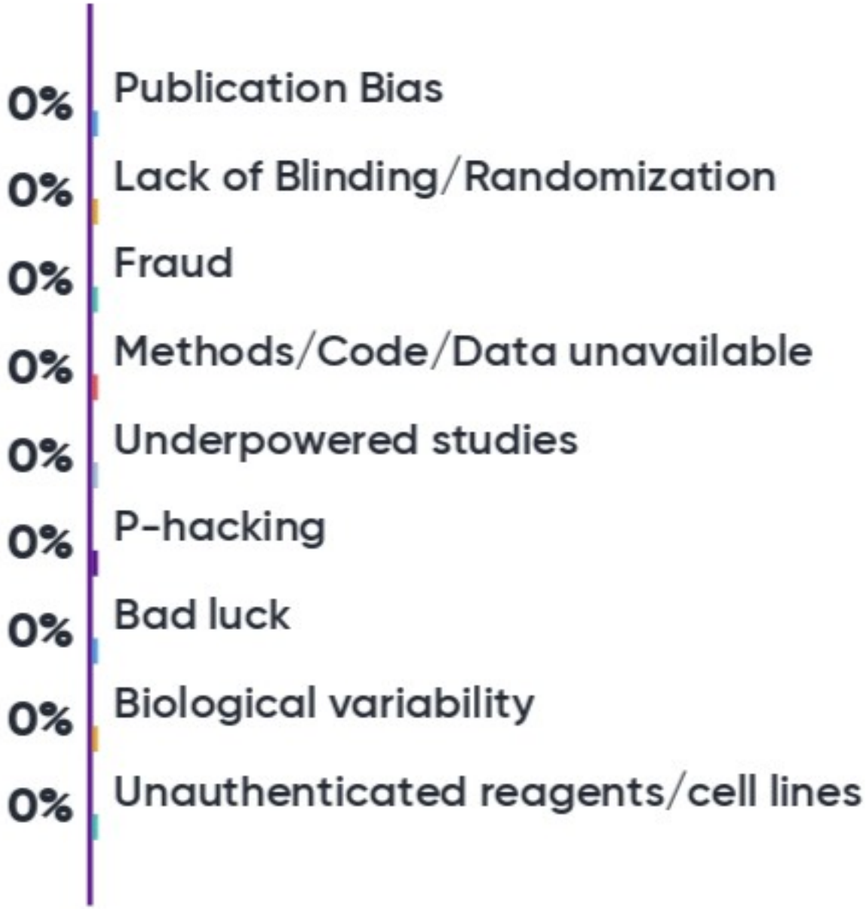
NYU HEALTH SCIENCES LIBRARY

Discuss with table: What measures have you or others taken to ensure replicability? Are there things that were not done that you would do now?

NYU HEALTH SCIENCES LIBRARY

Assign points based on how big a factor you think it is in the current reproducibility/replicability crisis

NYU HEALTH SCIENCES LIBRARY



0

NYU HEALTH SCIENCES LIBRARY

NIH

Element of Rigor	Section of Application	Criterion Score	Additional Review Consideration	Contribute to Overall Impact?
Scientific Premise	Research Strategy	Significance	NA	Yes
Scientific Rigor		Approach	NA	Yes
Consideration of Relevant Biological Variables Such as Sex		Approach	NA	Yes
Authentication of Key Biological and/or Chemical Resources	New Attachment	NA	Adequate or Inadequate	No

NIH wants four areas of clarification

Example from NIH grant apps

Male and female mice will be randomly allocated to experimental groups at age 3 months. At this age the accumulation of CUG repeat RNA, sequestration of MBNL1, splicing defects, and myotonia are fully developed. The compound will be administered at 3 doses (25%, 50%, and 100% of the MTD) for 4 weeks, compared to vehicle-treated controls. IP administration will be used unless biodistribution studies indicate a clear preference for the IV route. A group size of n = 10 (5 males, 5 females) will provide 90% power to detect a 22% reduction of the CUG repeat RNA in quadriceps muscle by qRT-PCR (ANOVA, α set at 0.05). The treatment assignment will be blinded to investigators who participate in drug administration and endpoint analyses. This laboratory has previous experience with randomized allocation and blinded analysis using this mouse model [refs]. Their results showed good reproducibility when replicated by investigators in the pharmaceutical industry [ref].

https://grants.nih.gov/reproducibility/index.htm

Scientific Premise

We will discuss premise more in the section on literature reviews, but one should clarify past research experience and evidence from literature.

Example from NIH grant apps

Male and female mice will be randomly allocated to experimental groups at **age 3 months. At this age the accumulation of CUG repeat RNA, sequestration of MBNL1, splicing defects, and myotonia are fully developed.** The compound will be administered at 3 doses (25%, 50%, and 100% of the MTD) for 4 weeks, compared to vehicle-treated controls. IP administration will be used unless biodistribution studies indicate a clear preference for the IV route. A group size of n = 10 (5 males, 5 females) will provide 90% power to detect a 22% reduction of the CUG repeat RNA in quadriceps muscle by qRT-PCR (ANOVA, α set at 0.05). The treatment assignment will be blinded to investigators who participate in drug administration and endpoint analyses. **This laboratory has previous experience with randomized allocation and blinded analysis using this mouse model [refs]. Their results showed good reproducibility when replicated by investigators in the pharmaceutical industry [ref].**

https://grants.nih.gov/reproducibility/index.htm

Scientific Rigor

- Robust, well-controlled and replicable experiments
- Clear steps taken to remove bias: blinding assessors, randomization happening, clear terms used

Example from NIH grant apps

Male and female mice will be **randomly allocated** to experimental groups at age 3 months. At this age the accumulation of CUG repeat RNA, sequestration of MBNL1, splicing defects, and myotonia are fully developed. The **compound will be administered at 3 doses (25%, 50%, and 100% of the MTD) for 4 weeks**, compared to **vehicle-treated controls**. IP administration will be used unless biodistribution studies indicate a clear preference for the IV route. A group size of **n = 10 (5 males, 5 females)** will provide **90% power** to detect a 22% reduction of the CUG repeat RNA in quadriceps muscle by **qRT-PCR (ANOVA, α set at 0.05)**. The treatment assignment will be **blinded to investigators** who participate in drug administration and endpoint analyses. This laboratory has previous experience with randomized allocation and blinded analysis using this mouse model [refs]. Their results showed good reproducibility when replicated by investigators in the pharmaceutical industry [ref].

https://grants.nih.gov/reproducibility/index.htm

Considerations of Relevant Biological Variables

- Should be factored into research designs, analyses, and reporting in vertebrate animal and human studies.
- Justifications for use of one sex if only using one, citing prior data and literature
- Cost and absence of known sex differences are inadequate justifications for not addressing sex.

Example from NIH grant apps

Male and female mice will be randomly allocated to experimental groups at age 3 months. At this age the accumulation of CUG repeat RNA, sequestration of MBNL1, splicing defects, and myotonia are fully developed. The compound will be administered at 3 doses (25%, 50%, and 100% of the MTD) for 4 weeks, compared to vehicle-treated controls. IP administration will be used unless biodistribution studies indicate a clear preference for the IV route. A group size of **n = 10 (5 males, 5 females)** will provide 90% power to detect a 22% reduction of the CUG repeat RNA in quadriceps muscle by qRT-PCR (ANOVA, α set at 0.05). The treatment assignment will be blinded to investigators who participate in drug administration and endpoint analyses. This laboratory has previous experience with randomized allocation and blinded analysis using this mouse model [refs]. Their results showed good reproducibility when replicated by investigators in the pharmaceutical industry [ref].

https://grants.nih.gov/reproducibility/index.htm

Submissions to NIH

- Premise, Rigor and Consideration of Relevant Biological Variables are part of the research strategy and scored
- Authentication Plan is an attachment, not scored
- Researchers determine what is key and plans should be based on scientific fields/disciplines

Review

What does it mean if a p value is 0.03?

NYU HEALTH SCIENCES LIBRARY

What are some ways to improve the transparency of your study when reporting?

NYU HEALTH SCIENCES LIBRARY

0

Why does the sex of your study subjects relate to reproducibility?

NYU HEALTH SCIENCES LIBRARY

Homework

Write a short essay (500 words maximum) on the use of $p < 0.05$ for statistical significance. Discuss whether you think this value should stay the same, change, or that we should abandon the use of statistical significance. Make an argument for why you think this will improve current issues around p-values.

|

Due in Brightspace by 9:00am 2/19/2020

?

Bibliography

1. Nowogrodzki A. Inequality in medicine. Nature. 2017;550:S18.
2. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. N Engl J Med. 1983 Dec 1;309(22):1358-61.
3. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. Neuroscience and biobehavioral reviews. 2011;35(3):565-72.
4. Clayton JA. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiol Behav. 2018 Apr 1;187:2-5. doi:10.1016/j.physbeh.2017.08.012. Epub 2017 Aug 17.
5. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015 Mar 13;13(3):e1002106. doi:10.1371/journal.pbio.1002106. eCollection 2015 Mar.
6. Bishop, D. (2018). Simulating data to gain insights into power and p-hacking. [online] Slideshare.net. Available at: https://www.slideshare.net/deevybishop/simulating-data-to-gain-insights-intopower-and-phacking.
7. National Academies of Sciences, Engineering, and Medicine. (2016). Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. Washington, DC: The National Academies Press.
8. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLOS ONE. 2009;4(11):e7824.
9. Avey MT, Moher D, Sullivan KJ, Fergusson D, Griffin G, et al. (2016) The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. PLOS ONE 11(11): e0166733.
10. On the reproducibility (https://peerj.com/articles/148/)
22. Chavalarias D, Wallach J, Li A, Ioannidis JA. Evolution of reporting p values in the biomedical literature, 1990-2015. JAMA. 2016;315(11):1141-8.
23. Warren, M. First analysis of 'pre-registered' studies shows sharp rise in null findings. Nature. 2018 Oct 18.
24. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018;2(1):6-10.
25. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. Nature Human Behaviour. 2018;2(3):168-71.
26. Ioannidis JA. The proposal to lower p value thresholds to .005. JAMA. 2018;319(14):1429-30.
27. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506(7487):150-2.
28. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. doi:10.1038/d41586-019-00857-9.
29. Promoting reproducibility with registered reports. Nature Human Behaviour. 2017;1:0034.
30. Kaplan RM, Irvin VL. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. PLoS One. 2015 Aug 5;10(8):e0132382. doi:10.1371/journal.pone.0132382. eCollection 2015.
31. Teytelman L. No more excuses for non-reproducible methods. Nature. 2018;560(7719):411.
32. Checklists work to improve science. Nature. 2018 April 18.