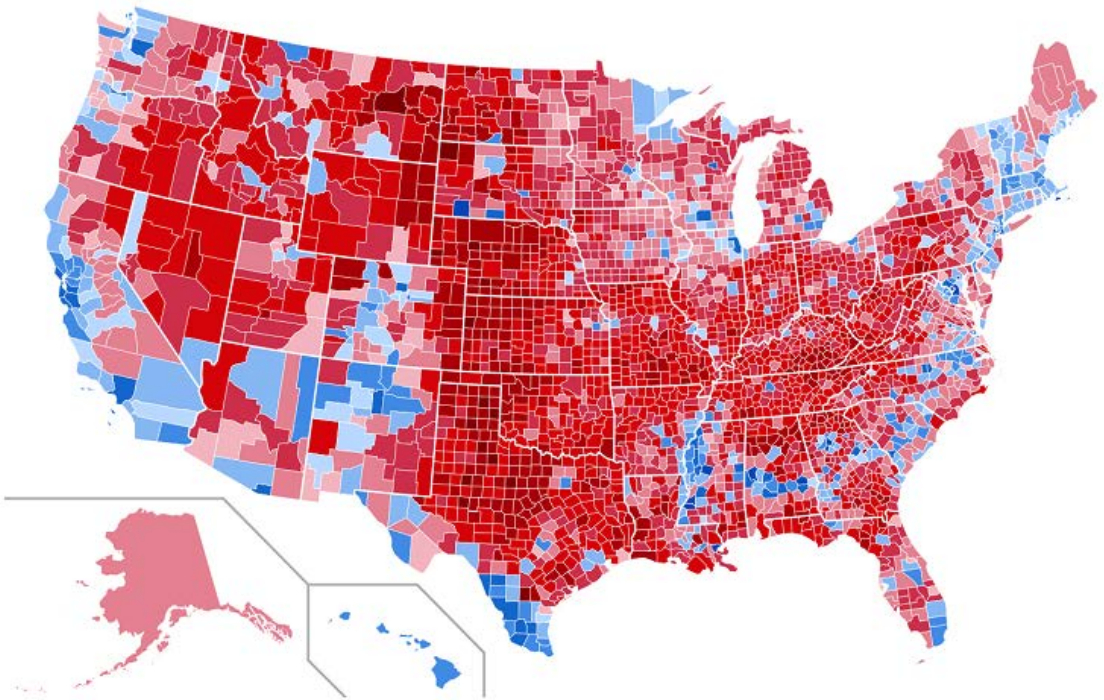
Perspectives on Enhancing Rigor and Transparency of Scientific Research

Michael S Lauer, MD
Deputy Director for Extramural Research, National Institutes of Health

Federal Demonstration Partnership Winter Meeting
Monday, January 8, 2018 (1 PM)
Hyatt Regency Capitol Hill, 400 NJ Avenue NW, Washington DC

Conflicts: None

What Do You Think About This?



Map created by Magog the Ogre via Wikimedia

"A study of the incidence of kidney cancer in the 3,141 counties of the US reveals a remarkable pattern. The counties in which the incidence of kidney cancer is lowest are mostly rural, sparsely populated, and located in traditionally [Red] states. What do you make of this?"

http://brilliantmaps.com/2016-county-election-map/

Kahneman D. Thinking Fast and Slow. FSG, 2011. Page 109

More Questions for You

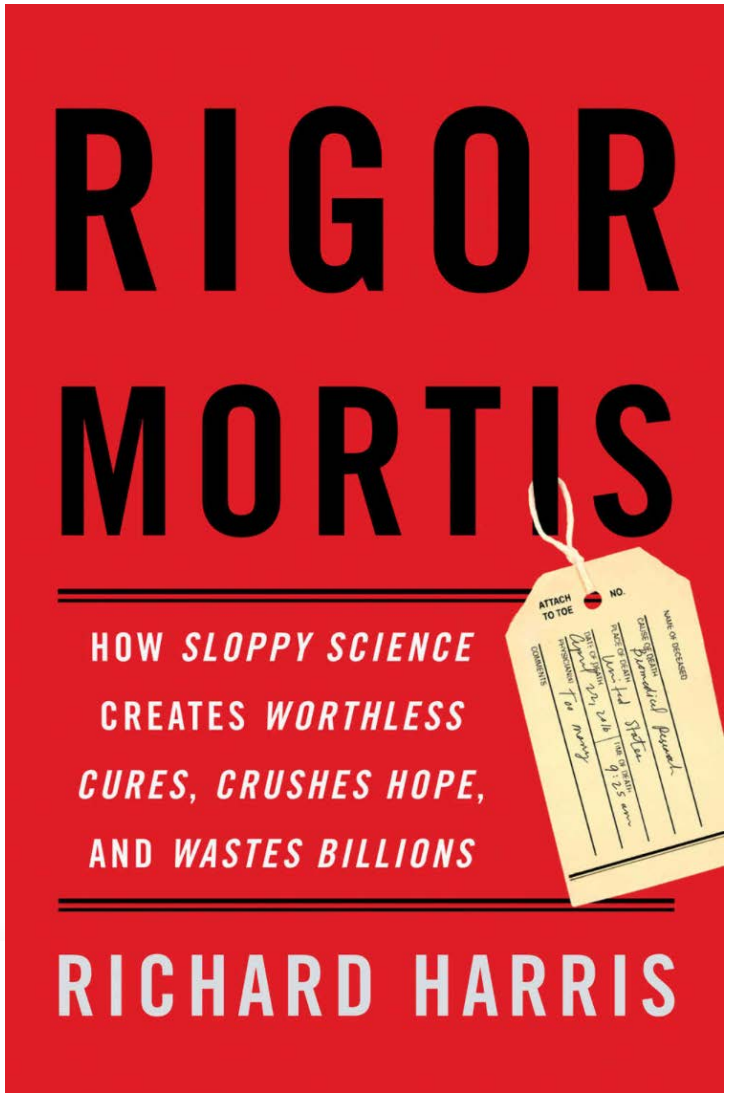
Real Worries!

Unreliable research

Trouble at the lab

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

"I SEE a train wreck looming," warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned



Problems:

- Animal models
- Cell lines
- Antibodies
- Poor study design
- Broken culture

http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Goes Further Back

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Lies, Damned Lies, and Medical Science

Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.

DAVID H. FREEDMAN | NOVEMBER 2010 ISSUE | TECHNOLOGY

PLoS Medicine 2005;2:e124
Atlantic Magazine, November 2010

PERSPECTIVE

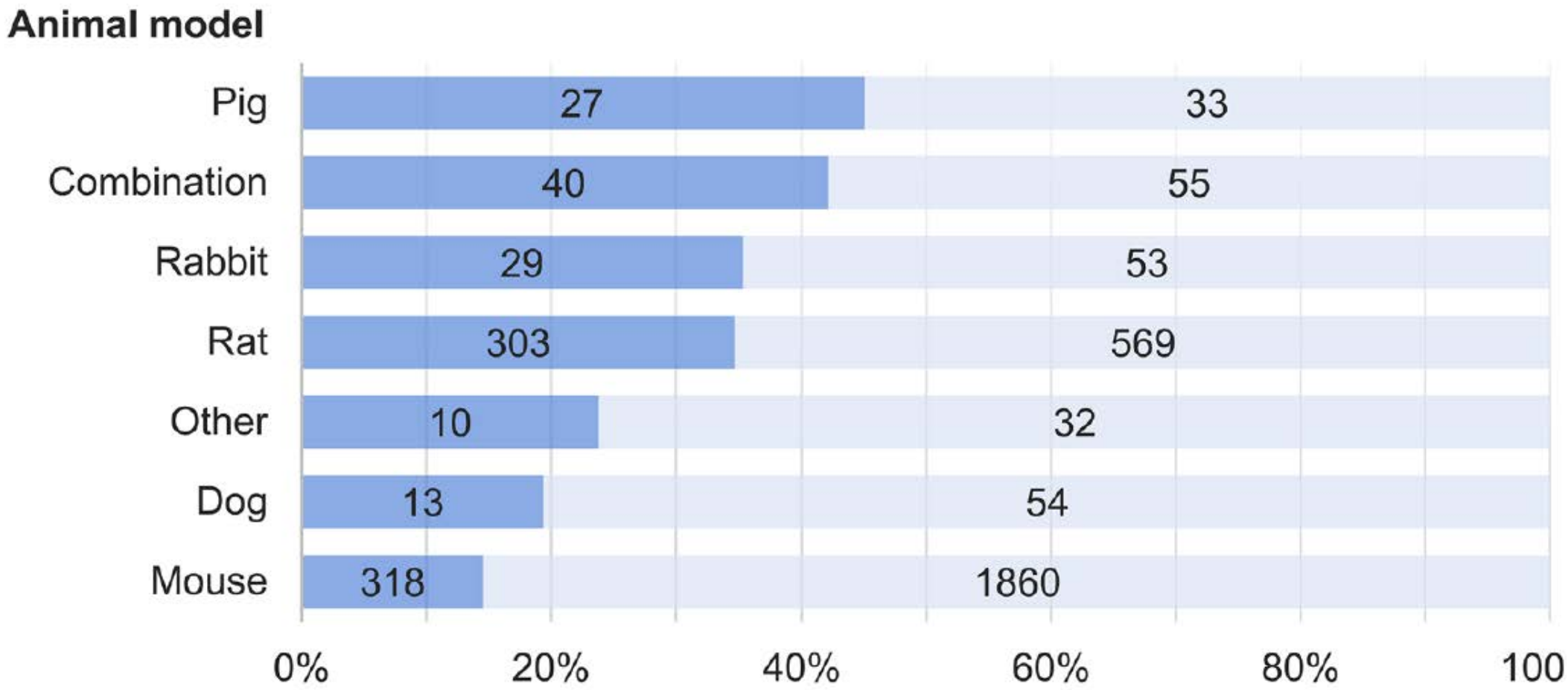
doi:10.1038/nature11556

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

"… At a mininum studies should report on **sample-size estimation**, whether and how animals were randomized, whether investigators were blind to the treatment, and the handling of data."

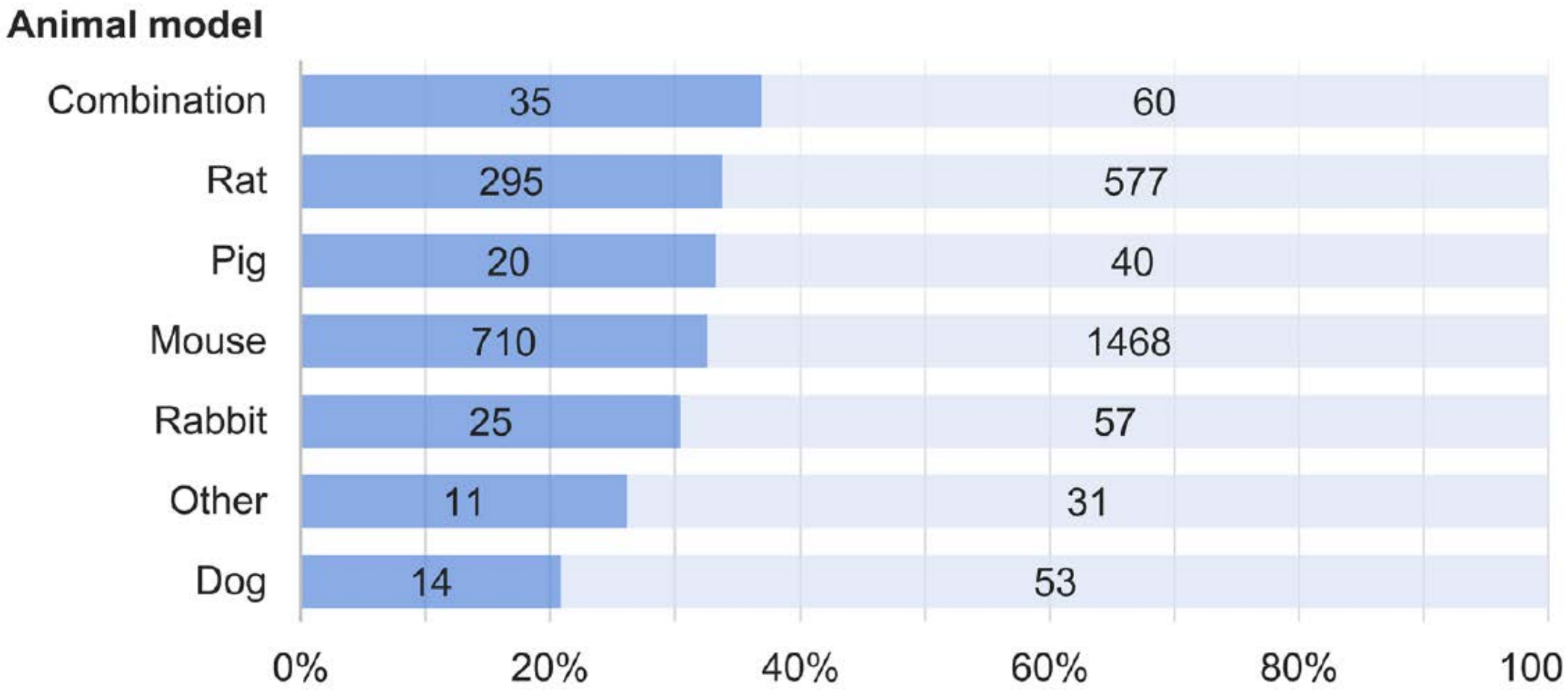
Landis SC, Silberberg S et al. Nature 2012;490:187-191

How Well Are We Doing? Randomization…



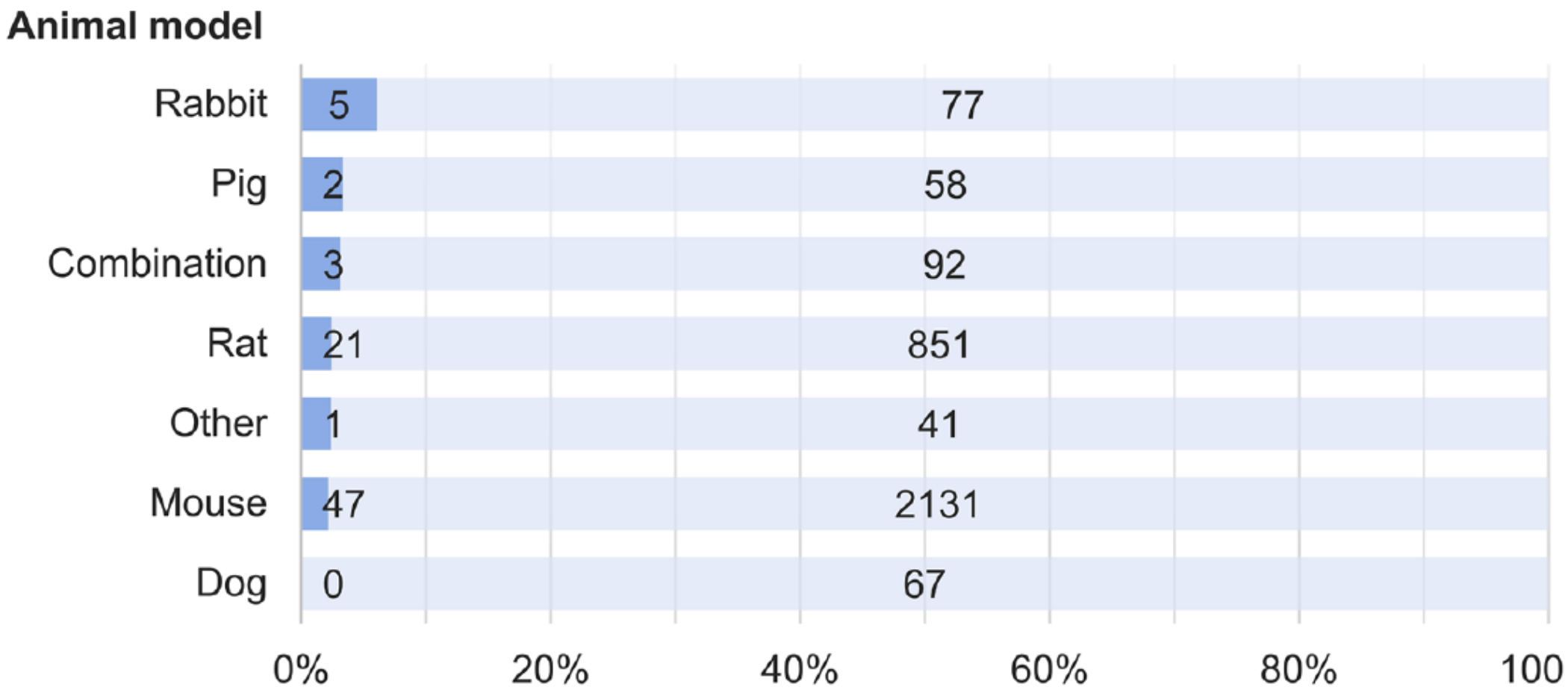
Ramirez FD, et al. Circulation Research. 2017;120:1916-26

Blinding …



Ramirez FD, et al. Circulation Research. 2017;120:1916-26

Sample Size Calculation …



Ramirez FD, et al. Circulation Research. 2017;120:1916-26

Inherent Problems

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

"For most study designs and settings, it is more likely for a research claim to be false than true."

- **Smaller studies**
- Smaller effect size
- Greater number of tested relationships
- Flexibility in designs and definitions
- Financial interests and fads

PLoS Medicine 2005;2:e124

THE LANCET

Vol 339 Saturday 27 June 1992 No 8809

ORIGINAL ARTICLES

Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2)

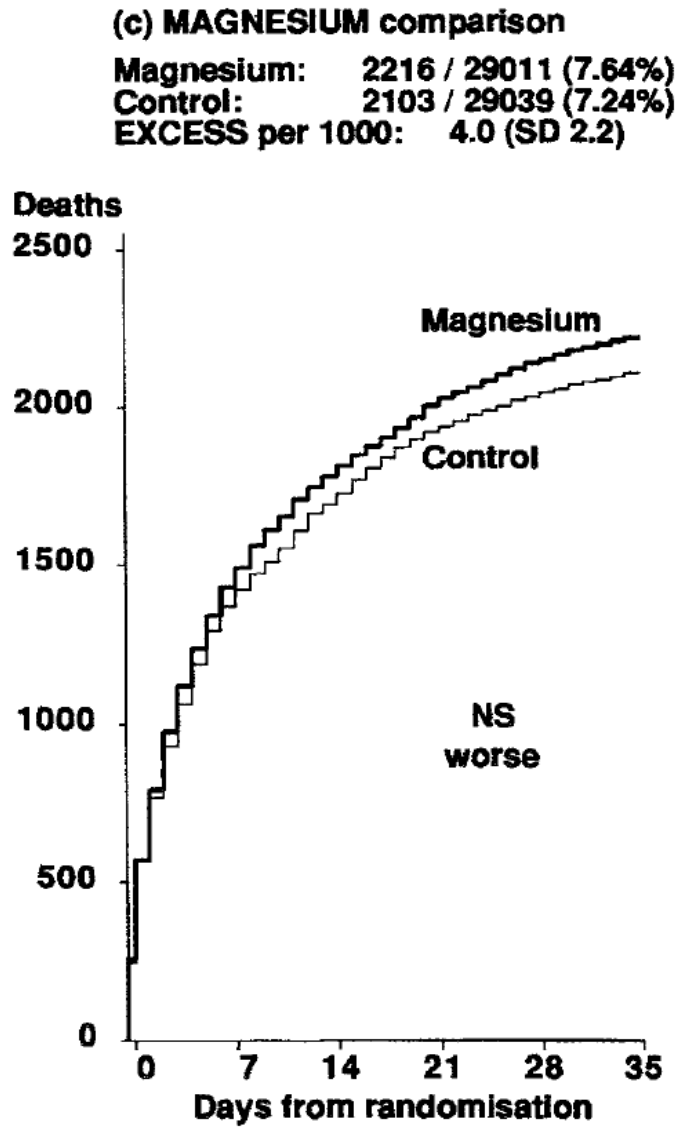
KENT L. WOODS SUSAN FLETCHER CHRISTINE ROFFE
YASSER HAIDER

The (Much) Bigger Study…

THE LANCET

ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction

*ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group**

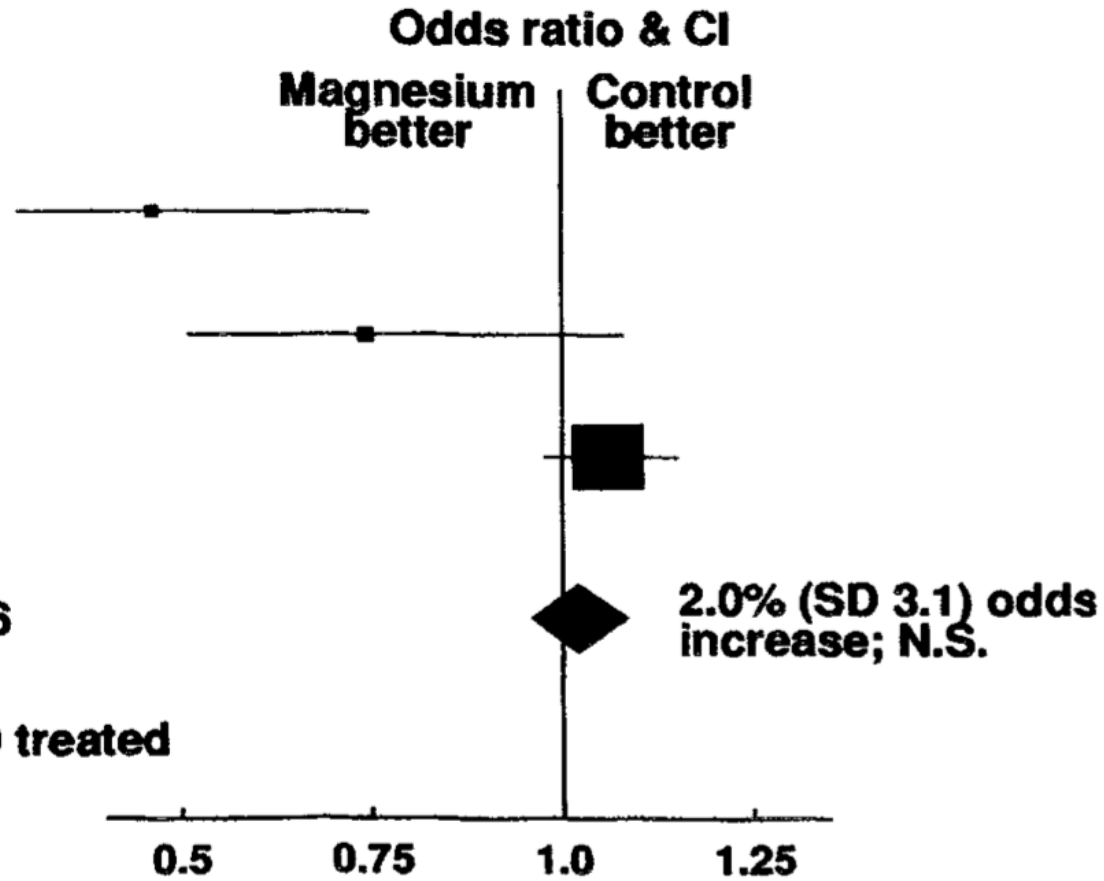


"Previously, eight **very small trials** of the intravenous infusion of magnesium … had collectively indicated a **mortality reduction of about one-half**."

Lancet 1995;345:669-85

What They Were Talking About …

Randomised comparison	Deaths/patients (% dead) Magnesium	Control
9 small trials	42/ 754 (5.6%)	86/ 740 (11.6%)
LIMIT-2	90/ 1159 (7.8%)	118/ 1157 (10.2%)
ISIS-4	2216/29011 (7.6%)	2103/29039 (7.2%)
ALL TRIALS	2348/30924 (7.59%)	2307/30936 (7.46%)

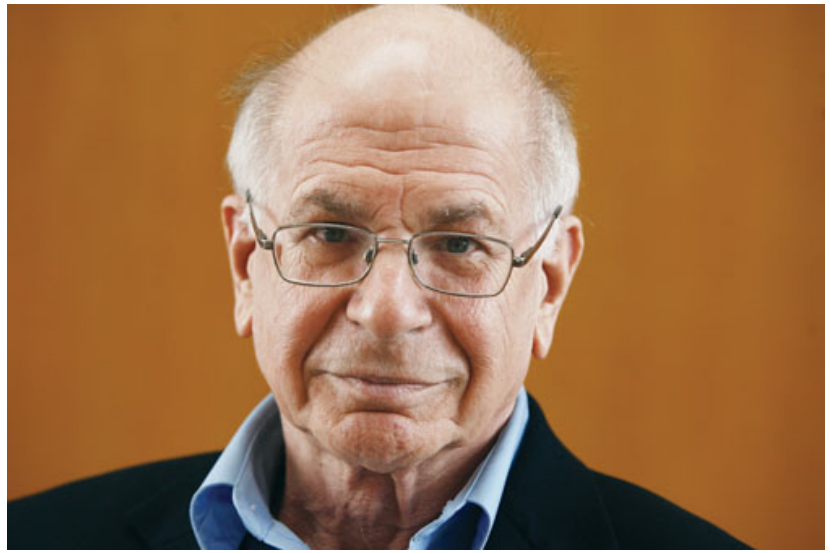


1.4 (SD 2.1) extra deaths per 1000 treated

Test for heterogeneity:

- between 9 small trials & 2 larger trials: $\chi^2_1 = 18.6$; $p<0.0001$
- between LIMIT-2 & ISIS-4: $\chi^2_1 = 5.7$; $p=0.02$

It's Fundamental: "Law of Small Numbers"

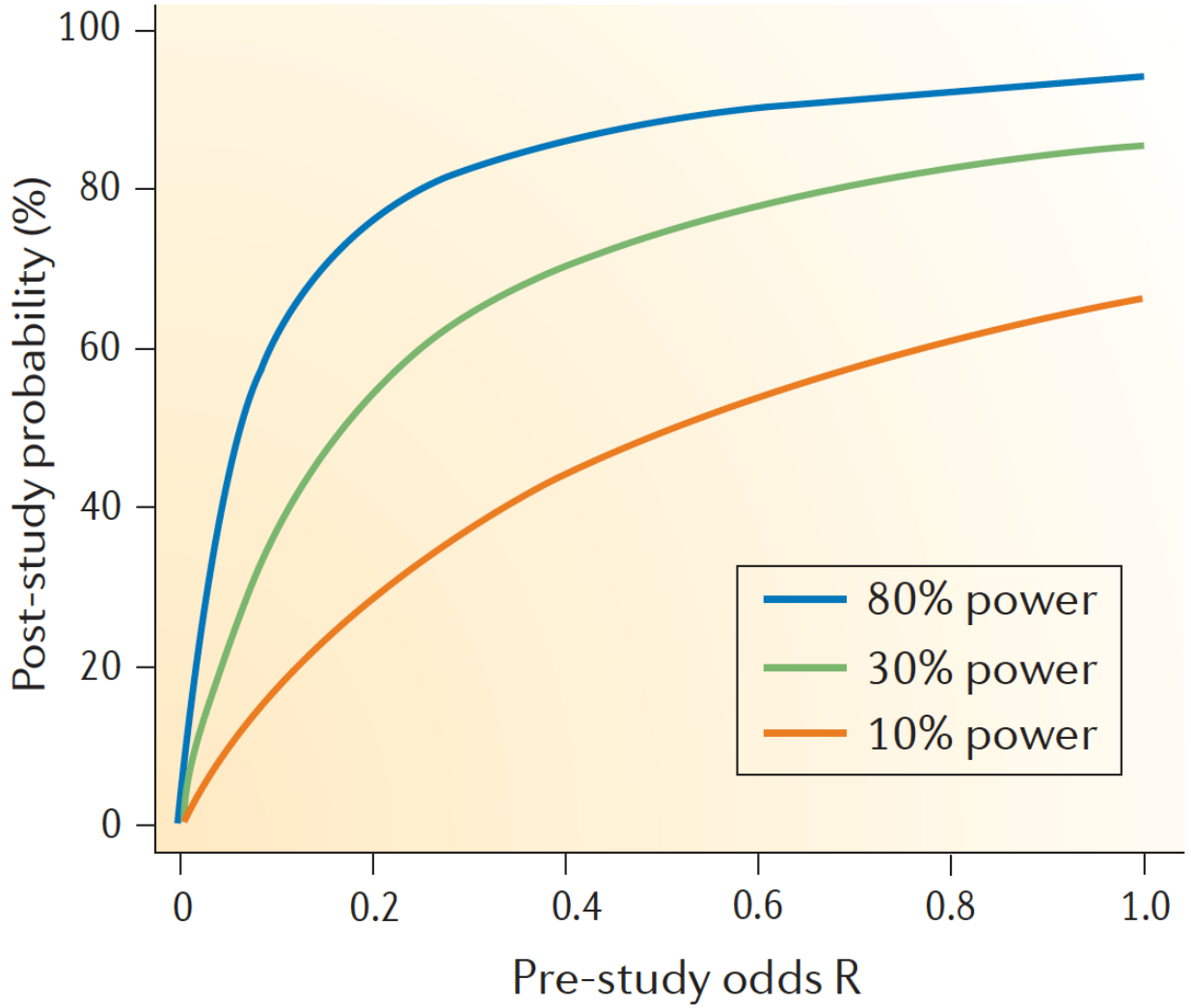


"Large samples are more precise than small samples. [This means that] small samples yield extreme results more often than large samples. The exaggerated faith in small samples is only one example of a more general illusion: a view of the world that is simpler than the data justify."

Daniel Kahneman

Kahneman D. Farrar, Straus, and Giroux. 2011

The “Power” of Having More “Power”



Button K, et al. Nature Review Neuroscience. 2013;14:365-76

“Power Failure”

ANALYSIS

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

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

	Total animals used	Required *N* per study		Typical *N* per study	
		80% power	95% power	Mean	Median
Water maze	420	134	220	22	20
Radial maze	514	68	112	24	20

“What is particularly striking is the inefficiency of a continued reliance on small sample sizes. … Low power has an ethical dimention – unreliable research is inefficient and wasterful. This applies to both human and animal research."

Button KS et al. Nature Reviews Neuroscience. 2013;14:365-76

Beginning to Be Recognized

USAF/GETTY

Experiments that use only a small number of animals are common, but might not give meaningful results.

MEDICAL RESEARCH

UK funders demand strong statistics for animal studies

Move addresses concerns that some experiments are not using enough animals.

BY DANIEL CRESSEY

Replace, refine, reduce: the 3 Rs of ethical animal research are widely accepted around the world. But now the message for animal experiments. Funding applicants must now show that their work will provide statistically robust results — not just explain how it is justified and set out the ethical implications — or risk having their grant application rejected. Sert, who works on experimental design at the National Centre for the Replacement, Refinement and Reduction (NC3Rs) of Animals in Research in London. "These animals are going to be wasted."

Cressey D. Nature 2015;520:271-2

Inherent Problems

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

"For most study designs and settings, it is more likely for a research claim to be false than true."

- Smaller studies
- **Smaller effect size**
- Greater number of tested relationships
- Flexibility in designs and definitions
- Financial interests and fads

PLoS Medicine 2005;2:e124

Quiz Again: What Do You Think About This?

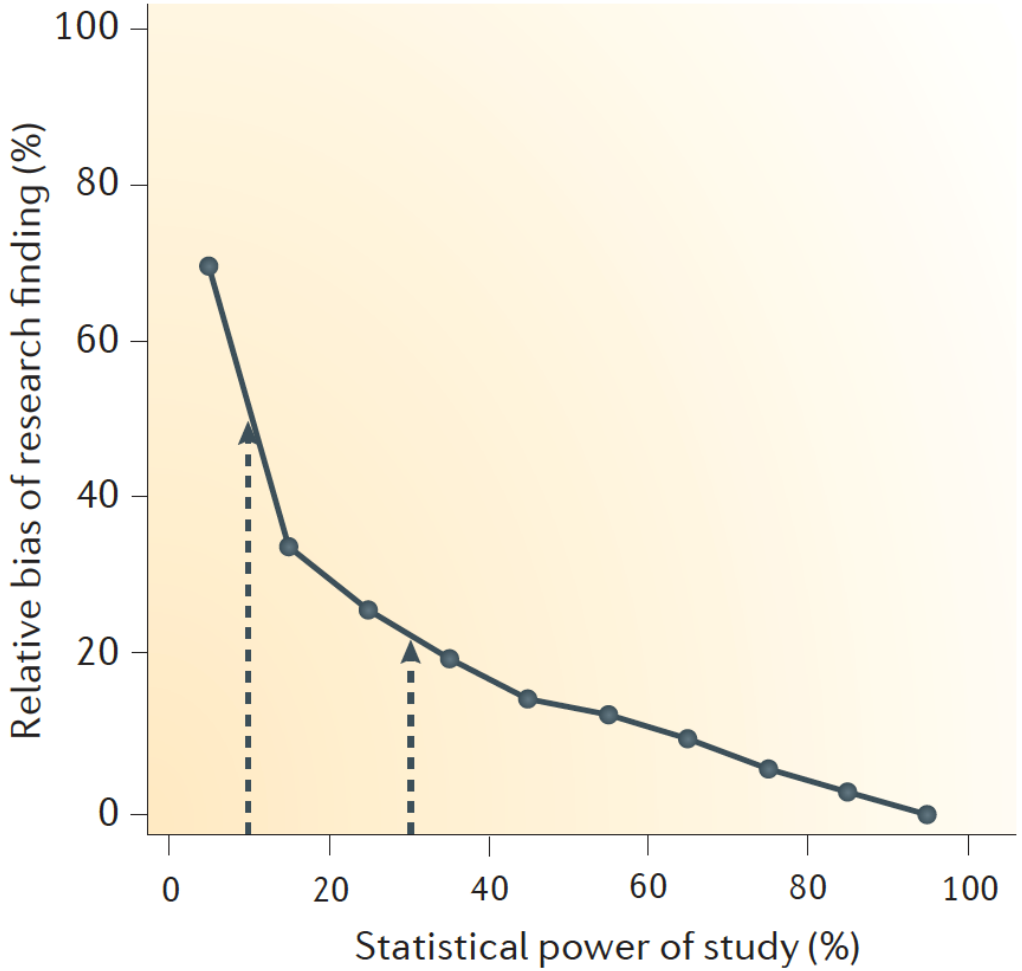
The New York Times

ART & DESIGN

Leonardo da Vinci Painting Sells for $450.3 Million, Shattering Auction Highs

By ROBIN POGREBIN and SCOTT REYBURN NOV. 15, 2017

The “Winner’s Curse”



“Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large. If the true effect is medium-sized, only those small studies that, **by chance**, estimate the effect to be large will pass the threshold … Research findings of small studies are biased in favor of inflated effects.”

Button K, et al. Nature Review Neuroscience. 2013;14:365-76

Inherent Problems

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

"For most study designs and settings, it is more likely for a research claim to be false than true."

- Smaller studies
- Smaller effect size
- **Greater number of tested relationships**
- **Flexibility in designs and definitions**
- Financial interests and fads

PLoS Medicine 2005;2:e124

Another Perspective: LOTS of Questions

FiveThirtyEight

Politics Sports **Science & Health** Economics Culture Catch up: Trump's inauguration

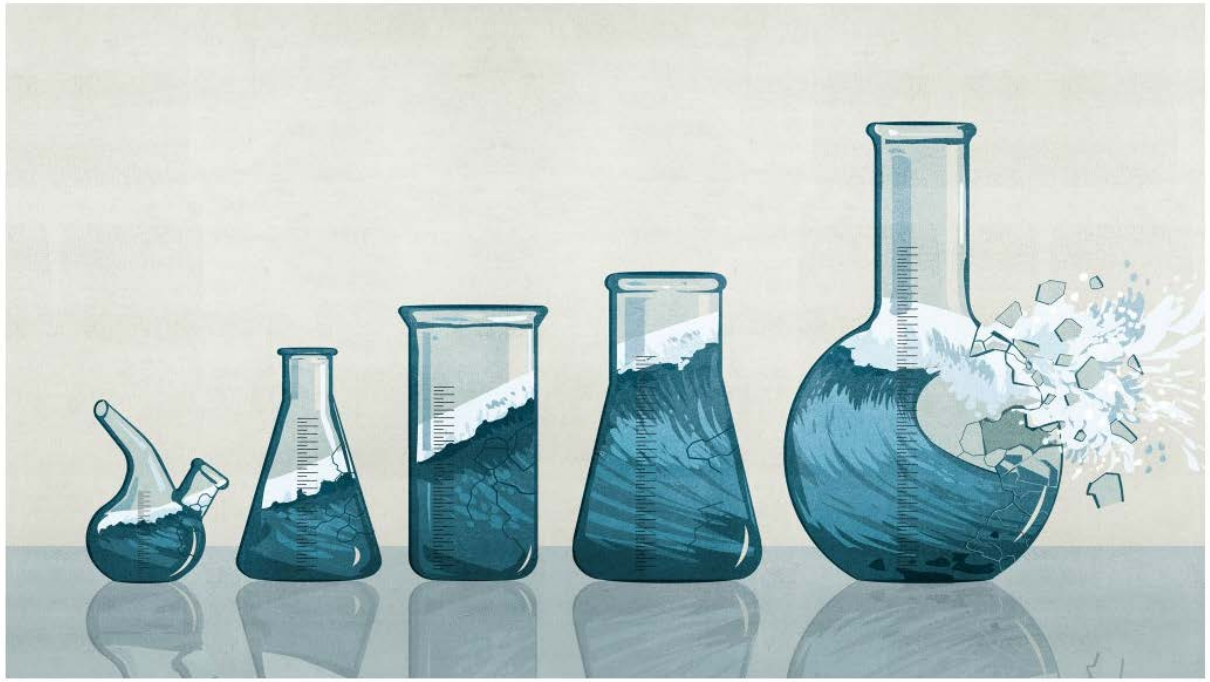


ILLUSTRATION BY SHOUT

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

By Christie Aschwanden
Filed under Scientific Method
Published Aug 19, 2015

https://fivethirtyeight.com/features/science-isnt-broken/#part1
https://projects.fivethirtyeight.com/p-hacking/

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [x] Senators
- [x] Representatives

How do you want to measure economic performance?

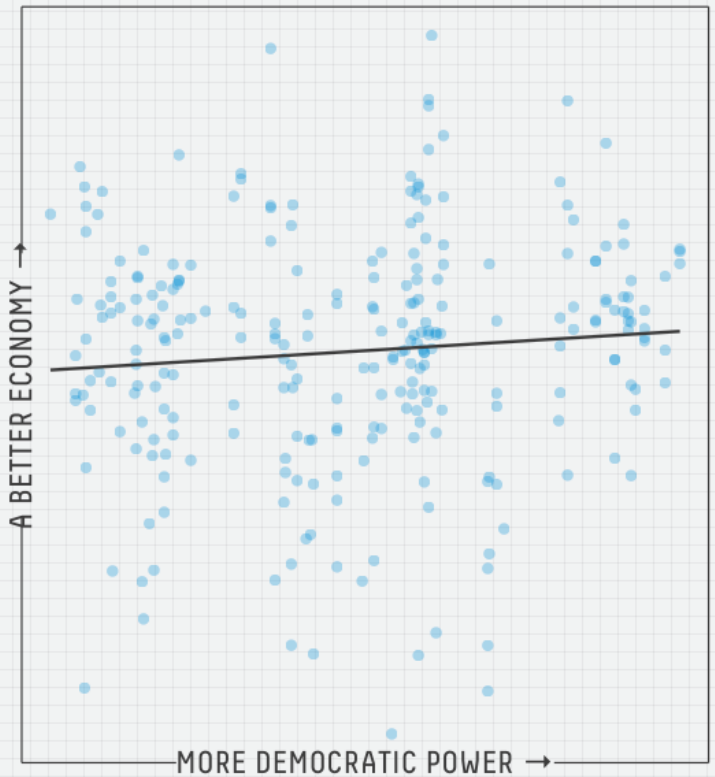
- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

Other options

- [x] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

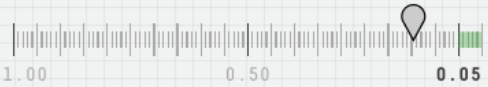
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Unpublishable

With a p-value of **0.15**, your findings are not statistically significant. Try defining your terms differently.

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

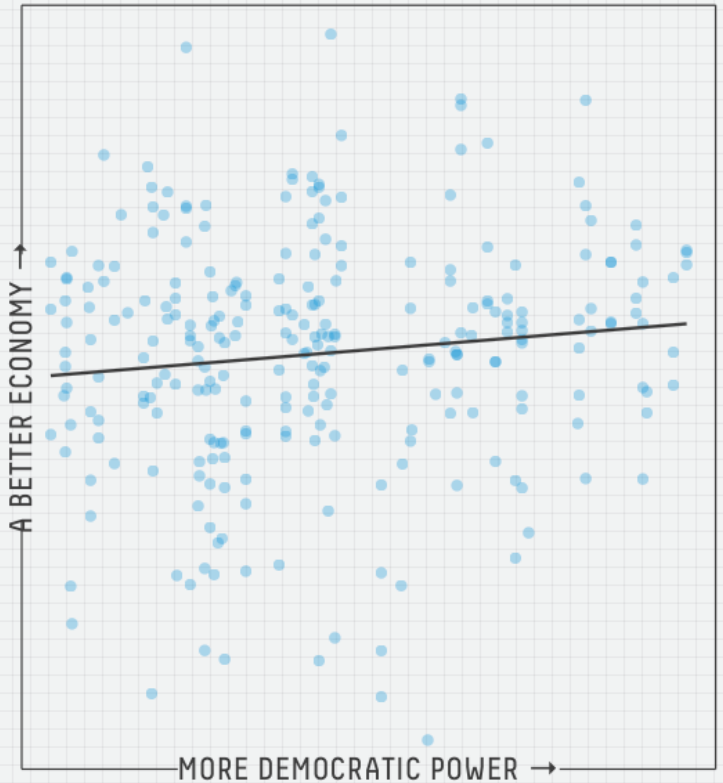
- [ ] **Employment**
- [x] **Inflation**
- [x] **GDP**
- [x] **Stock prices**

Other options

- [x] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [ ] Senators
- [x] Representatives

How do you want to measure economic performance?

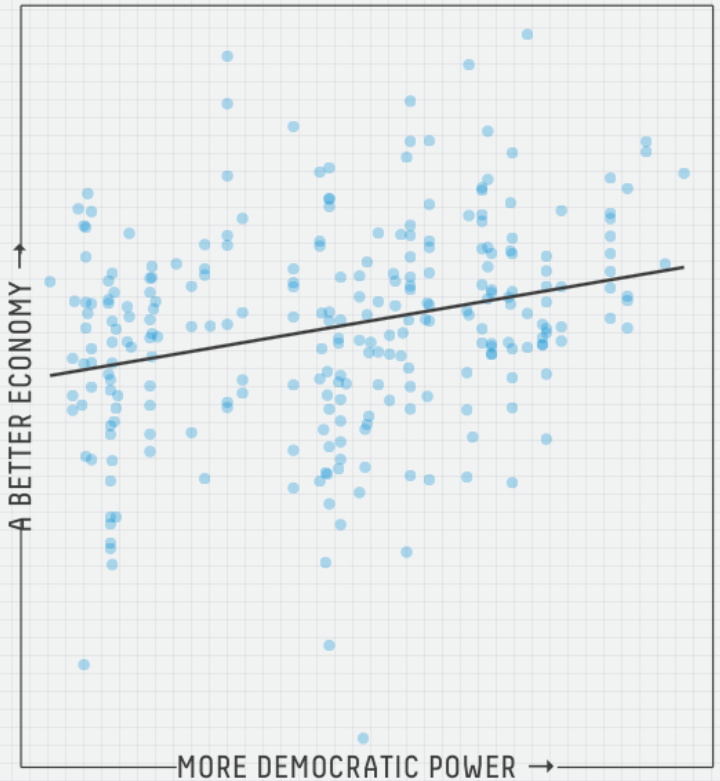
- [ ] Employment
- [ ] Inflation
- [x] GDP
- [x] Stock prices

Other options

- [x] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **positive** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Elsewhere

JOINT MATHEMATICS MEETINGS
AMERICAN MATHEMATICAL SOCIETY
MATHEMATICAL ASSOCIATION OF AMERICA

2017 JOINT MATHEMATICS MEETINGS

Largest Mathematics Meeting in the World

JAN 4-7 (WED - SAT), 2017 | HYATT REGENCY ATLANTA AND MARRIOTT ATLANTA MARQUIS

Courtesy of Linda A. Cicero, Stanford University

LECTURE IV

Statistical Proof and the Problem of Irreproducibility

Friday, January 6, 2017, Starting at 4:00 p.m. Imperial Ballroom A, Marquis Level, Marriott Marquis Atlanta

Susan Holmes, Stanford University

Data currently generated in the fields of ecology, medicine, climate science and neuroscience often contain tens of thousands of measured variables. Statistical analyses can result in publications whose results are irreproducible.

http://jointmathematicsmeetings.org/meetings/national/jmm2017/2180_invspeakers#holm2

Thanks to Dr. Jonathan Rosenberg

NIH National Institutes of Health
Office of Extramural Research

comment

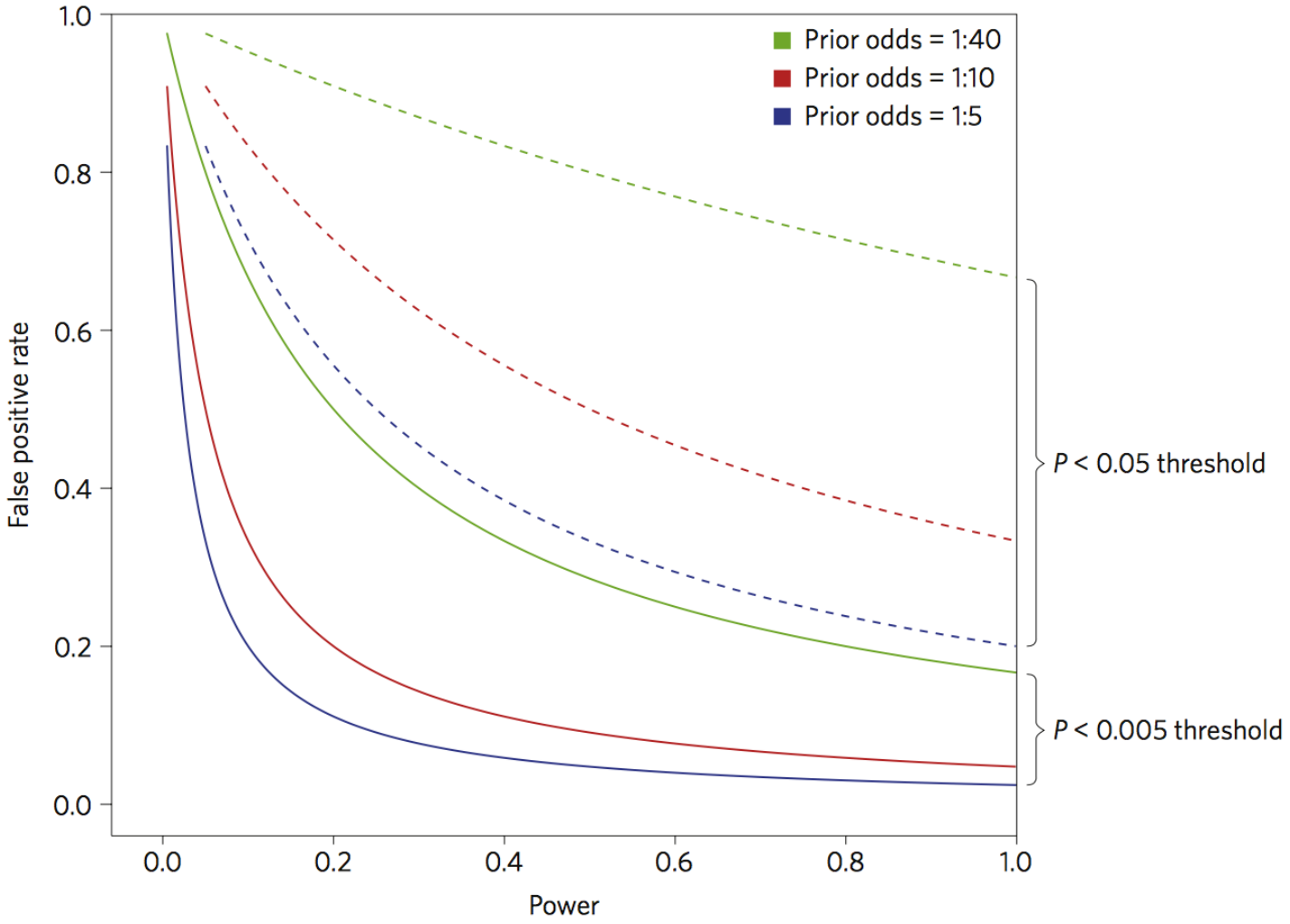
Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

"There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: **statistical standards of evidence for claiming new discoveries in many fields of science are simply too low.** Associating statistically significant findings with $P < 0.05$ results in a **high rate of false positives** even in the absence ..."

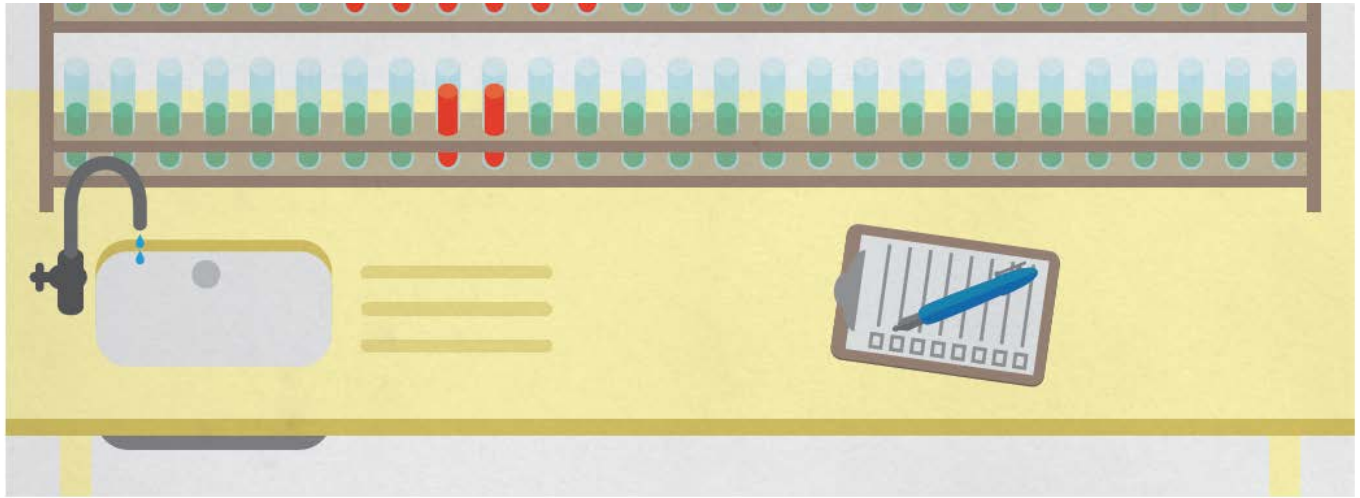
Nature Human Behavior 2017: https://www.nature.com/articles/s41562-017-0189-z.pdf

Putting It All Together …



"A much larger pool of scientists **are now asking a much larger number of questions**, possibly with much **lower prior odds** of success … Reducing the *P* value threshold for claims of new discoveries to 0.005 is an actionable step that will immediately improve reproducibility."

Nature Human Behavior 2017: https://www.nature.com/articles/s41562-017-0189-z.pdf



NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

Nature. 2014;505:612-13

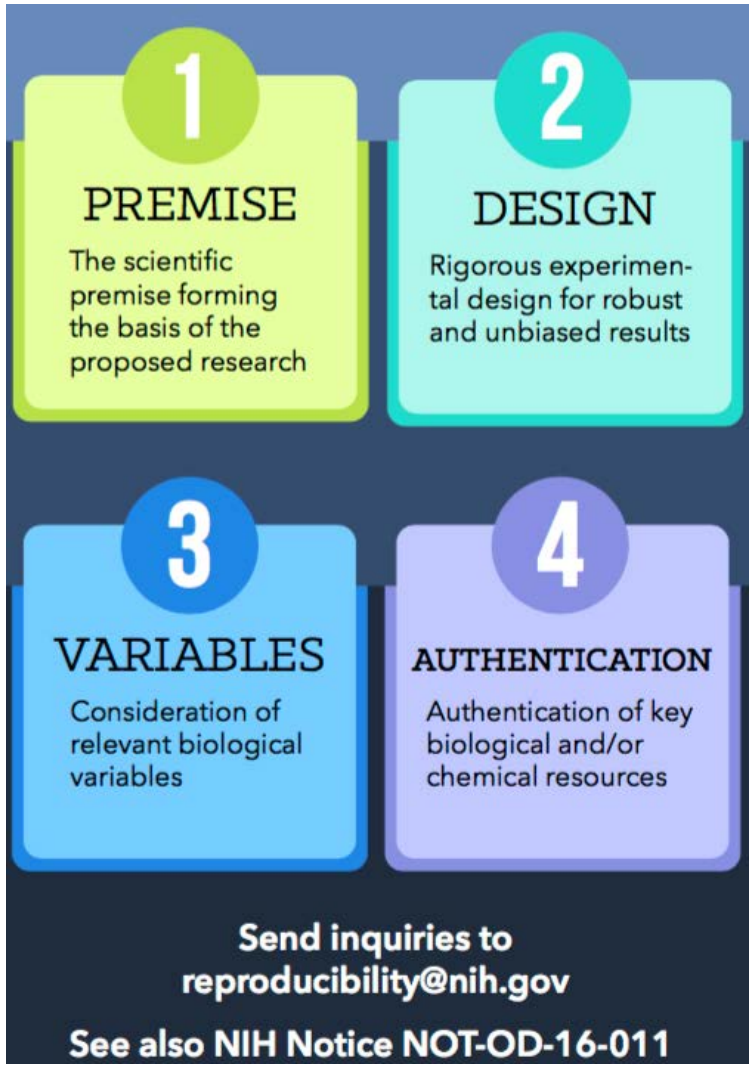
NIH National Institutes of Health
Office of Extramural Research | Grants & Funding
NIH's Central Resource for Grants and Funding Information

Rigor and Reproducibility

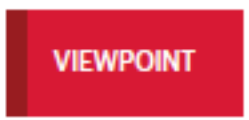
Scientific rigor and transparency in conducting biomedical resear
outcomes. The information provided on this website is designed
NIH grant applications and progress reports.

On This Page:

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



Special Steps with Clinical Trials



Toward a New Era of Trust and Transparency in Clinical Trials

Kathy L. Hudson, PhD
National Institutes of Health, Bethesda, Maryland.

Michael S. Lauer, MD
National Institutes of Health, Bethesda, Maryland.

Francis S. Collins, MD, PhD
National Institutes of Health, Bethesda, Maryland.

- Dedicated FOAs
- Special review criteria
- GCP Training
- Single IRB
- **Required registration, reporting**
- NIH-wide oversight system

JAMA 2016 (online pub September 16, 2016)

Many Stakeholders

EDITORIAL

Journals unite for reproducibility

Reproducibility, rigor, transparency, and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not necessarily make it right, and just because it is not reproducible does not necessarily make it wrong. A transparent and rigorous approach, however, can almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the resulting data.

It was with the goal of

menters were blind to the conduct of the experiment, how the sample size was determined, and what criteria were used to include or exclude any data. Journals should recommend the deposition of data in public repositories where available and link data bidirectionally to the published paper. Journals should strongly encourage, as appropriate, that all materials used in the experiment be shared with those who wish to replicate the experiment. Once a journal publishes a paper, it assumes the obligation to consider publication of a refutation of that paper, subject to its usual standards of quality.

The more open-ended portion of the guidelines suggests

Marcia McNutt
Editor-in-Chief
Science *Journals*

Enhancing Research Reproducibility:

Recommendations from the Federation of American Societies for Experimental Biology

FASEB
Federation of American Societies for Experimental Biology

Science. 2014;346:679

Appreciate Help With Communications

Science 2014;346:1452-3
https://www.gbsi.org/event/asilomar/
FASEBJ. 2016;30:519-24

Section 2039 requires the NIH Director to convene a working group under the ACD to develop and issue recommendations through the ACD for a formal policy, which may incorporate or be informed by relevant existing and ongoing activities, **to enhance rigor and reproducibility of scientific research funded by NIH**.

Concluding Thoughts: Longstanding, Core Issues

- Small numbers
- Regression to the mean
- Prior probability
- Multiple comparisons
- Misunderstood P-values
- Absence of transparency, full reporting
- We appreciate your help!