Beyond the Bench: Bridging Biostatistics and Biomedical Research for Reproducibility and Translation

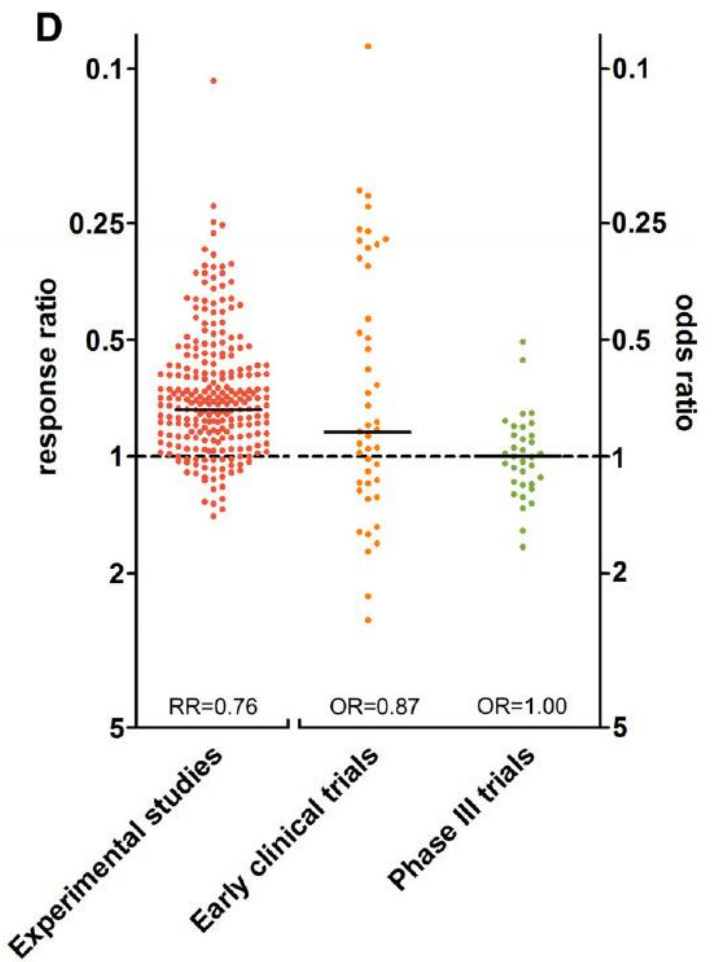
Non-Clinical Statistics Conference, Wiesbaden, 26.9.2024

Slide download: http://bit.ly/dirnaglncs

bitly

BIH QUEST
Center for Responsible Research

Total translational attrition in stroke research



Why Most Acute Stroke Studies Are Positive in Animals but Not in Patients: A Systematic Comparison of Preclinical, Early Phase, and Phase 3 Clinical Trials of Neuroprotective Agents

ANN NEUROL 2020;87:40–51

https://onlinelibrary.wiley.com/doi/full/10.1002/ana.25643

The SAINT-II experience: 5 billion US$ lost....

The NEW ENGLAND JOURNAL of MEDICINE

N Engl J Med 2007;357:562-71.

ORIGINAL ARTICLE

NXY-059 for the Treatment of Acute Ischemic Stroke

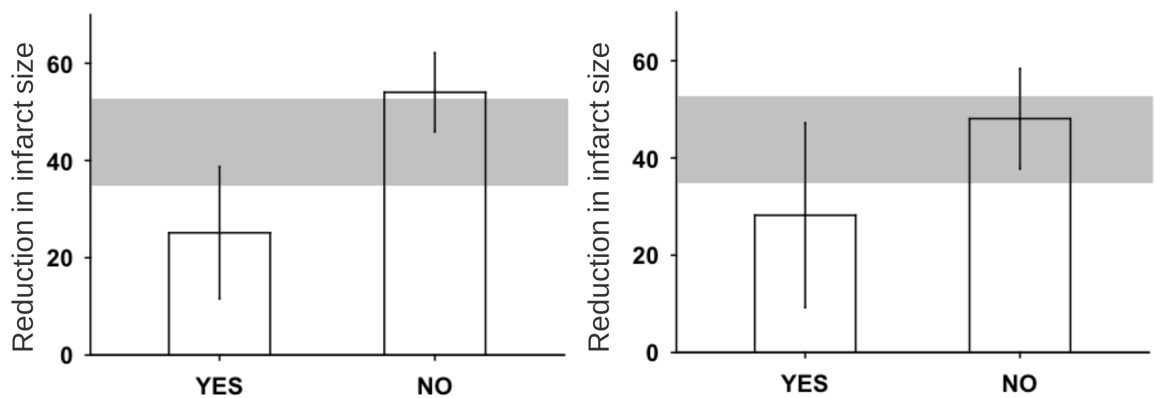
BIH QUEST
Center for Responsible Research

Selection and performance bias:
False positives and inflated effect sizes

Evidence for the Efficacy of NXY-059 in Experimental Focal Cerebral Ischaemia Is Confounded by Study Quality

Malcolm R. Macleod, PhD, FRCP; H. Bart van der Worp, MD, PhD; Emily S. Sena, BSc;
David W. Howells, PhD; Ulrich Dirnagl, MD, PhD; Geoffrey A. Donnan, MD, FRACP

Stroke models (NXY-095)



Blinded conduct of experiment

Blinded assessment of outcome

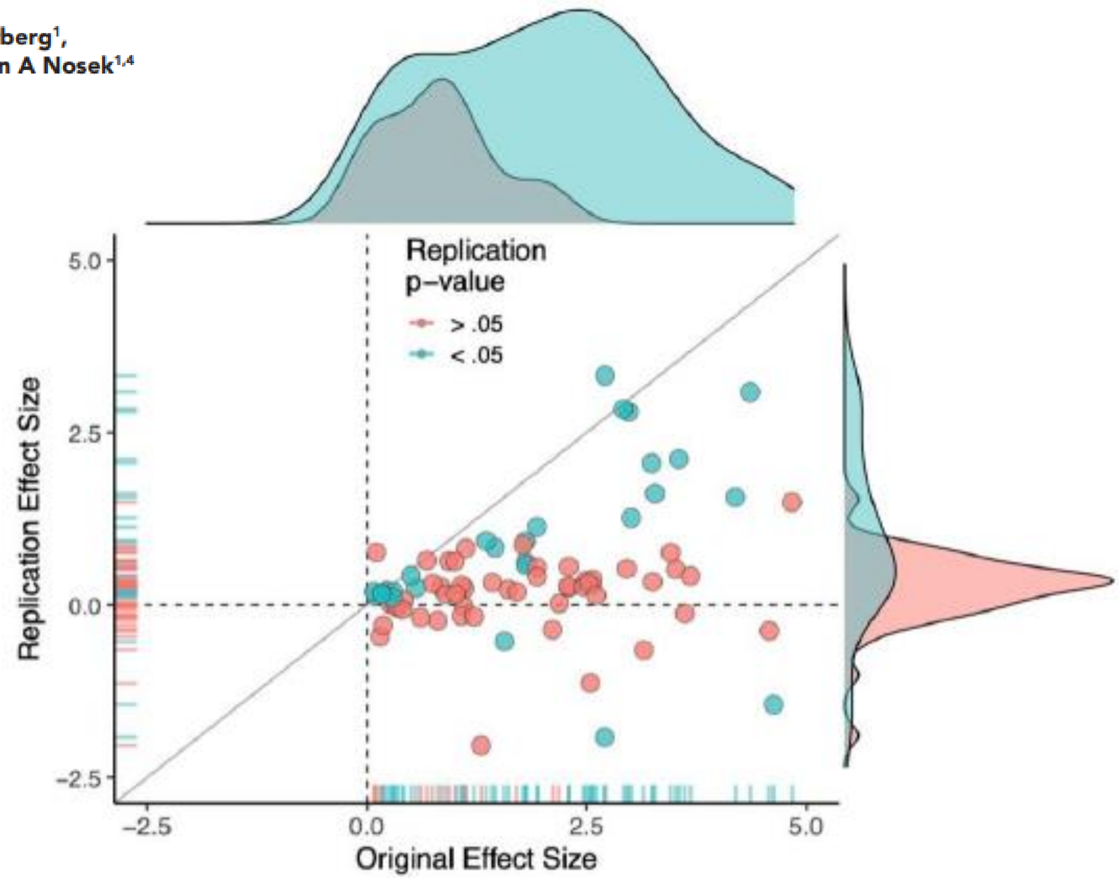
> 30 studies > 500 animals

https://doi.org/10.1161/STROKEAHA.108.515957

Reproducibility ‚crisis' exposed

Investigating the replicability of preclinical cancer biology

Timothy M Errington[1*], Maya Mathur[2], Courtney K Soderberg[1], Alexandria Denis[1†], Nicole Perfito[1‡], Elizabeth Iorns[3], Brian A Nosek[1,4]



Errington et al. https://elifesciences.org/articles/71601

Important reasons for non-reproducibility and translational attrition I will (today) NOT talk about

(Patho) Biological complexity

Low internal validity
(selection/performance/detection/attrition/... bias)

Humans are not 70 kg mice

Low external and construct validity

Publication bias

BIH QUEST
Center for Responsible Research

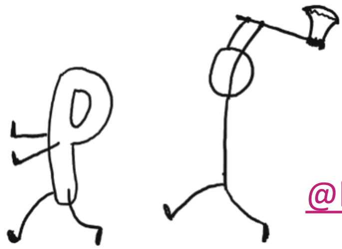
I will talk about:
Which role did bad statistics play in this mess?
Which role can better statistics play to get out of it?

- Small sample sizes, lack of statistical power, sample size sambas
- Inflation of effect sizes
- Statistical threshold for claiming a discovery too low ($p < 0.05$)
- *p*-hacking, uncorrected multiple comparisons, HARKING
- Lack of understandig basic statistical concepts (‚statistical significance‘, ‚prior probability‘,‚regression to the mean‘, etc.)
- Garden of the forking paths

… collectively leading to an inflation of false positives, false negatives, and effect sizes

BIH QUEST
Center for Responsible Research

p-Hacking



@kareem_carr

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Check for updates

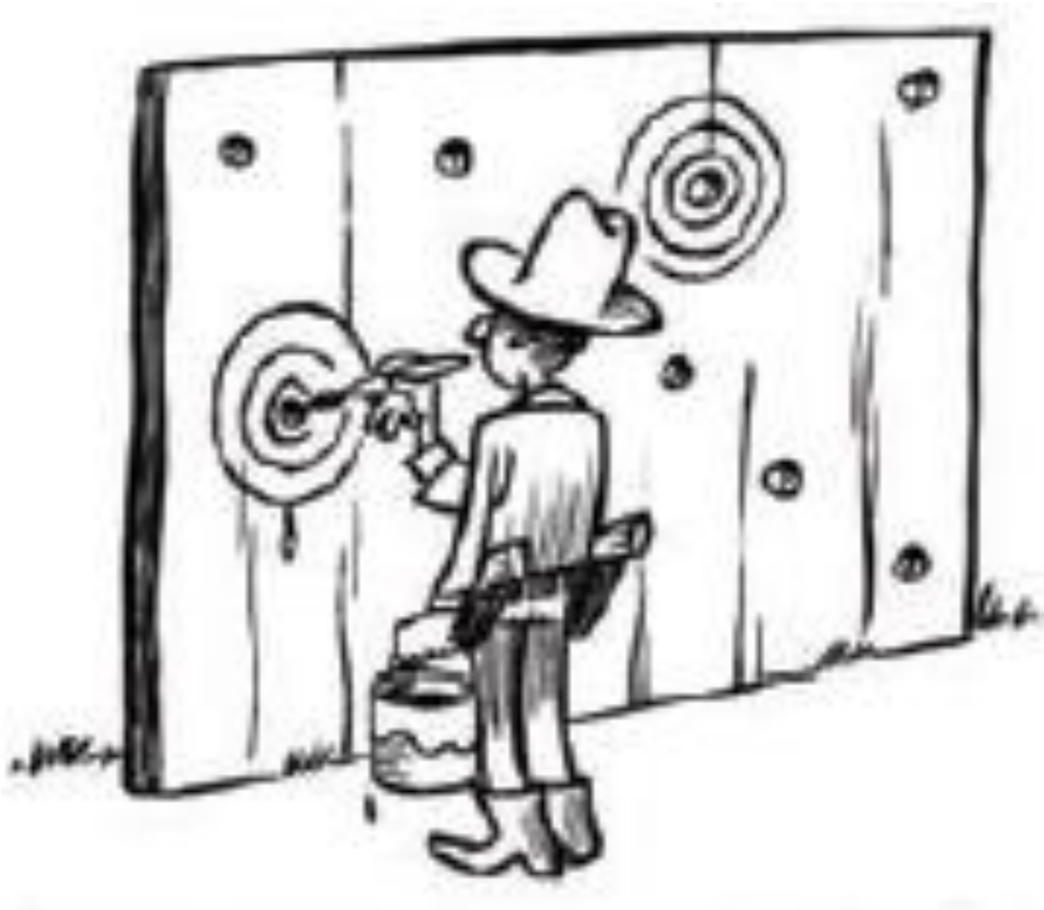
Big little lies: a compendium and simulation of *p*-hacking strategies

Angelika M. Stefan[1,2] and Felix D. Schönbrodt[3]

https://royalsocietypublishing.org/doi/full/10.1098/rsos.220346

BIH QUEST
Center for Responsible Research

HARKING: *Hypothesizing after the results are known*



An Agenda for Purely Confirmatory Research

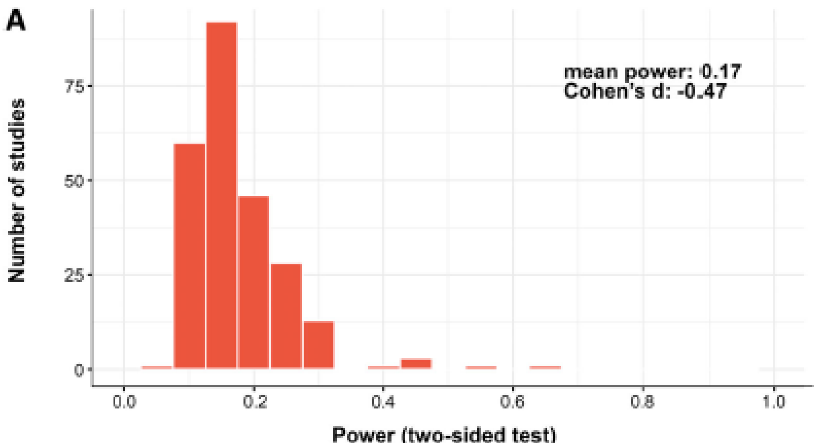
Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit
University of Amsterdam, The Netherlands

Perspectives on Psychological Science
7(6) 632–638

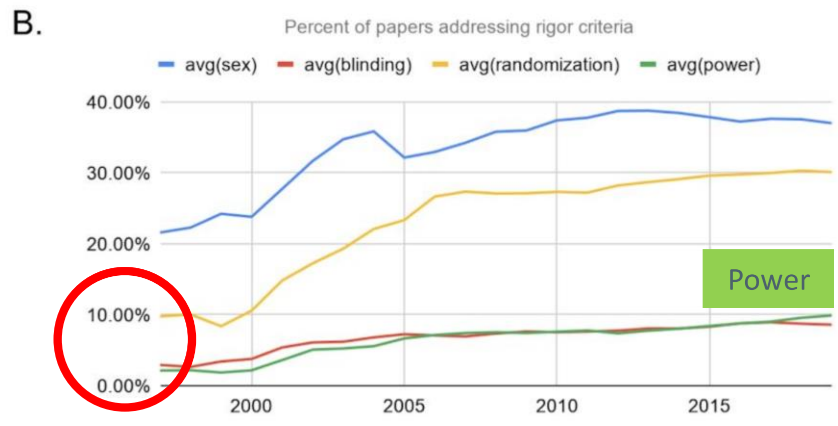
DOI: 10.1177/1745691612463078

https://doi.org/10.1177/1745691612463078

Exceedingly low sample sizes and statistical power in preclinical research



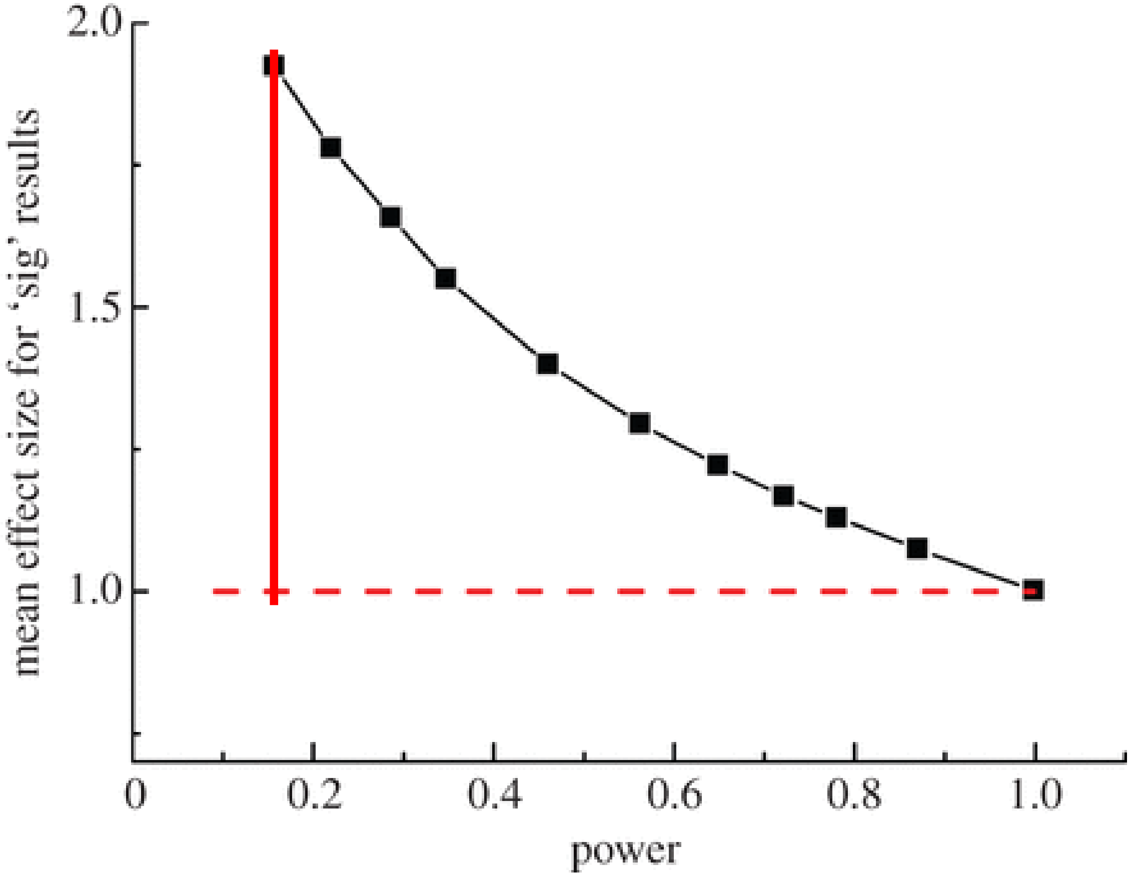
https://onlinelibrary.wiley.com/doi/full/10.1002/ana.25643



https://www.sciencedirect.com/science/article/pii/S2589004220308907

BIH QUEST
Center for Responsible Research

“Low sample size bias“ leads to false negatives, false positives, AND effect size inflation (*Winner’s curse*)



http://rsos.royalsocietypublishing.org/content/1/3/140216

BIH QUEST
Center for Responsible Research

Sample size samba

‘Retrofitting of the parameter estimates (in particular, the treatment effect worthy of detection) to the available participants’

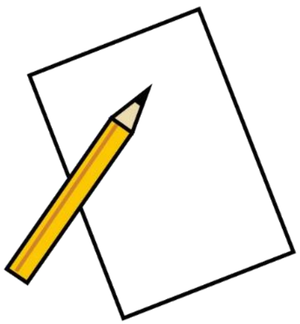
Schulz & Grimes; Sample size calculations in randomised trials: mandatory and mystical
The Lancet, 365, 1348-1353 (2005) https://doi.org/10.1016/S0140-6736(05)61034-3

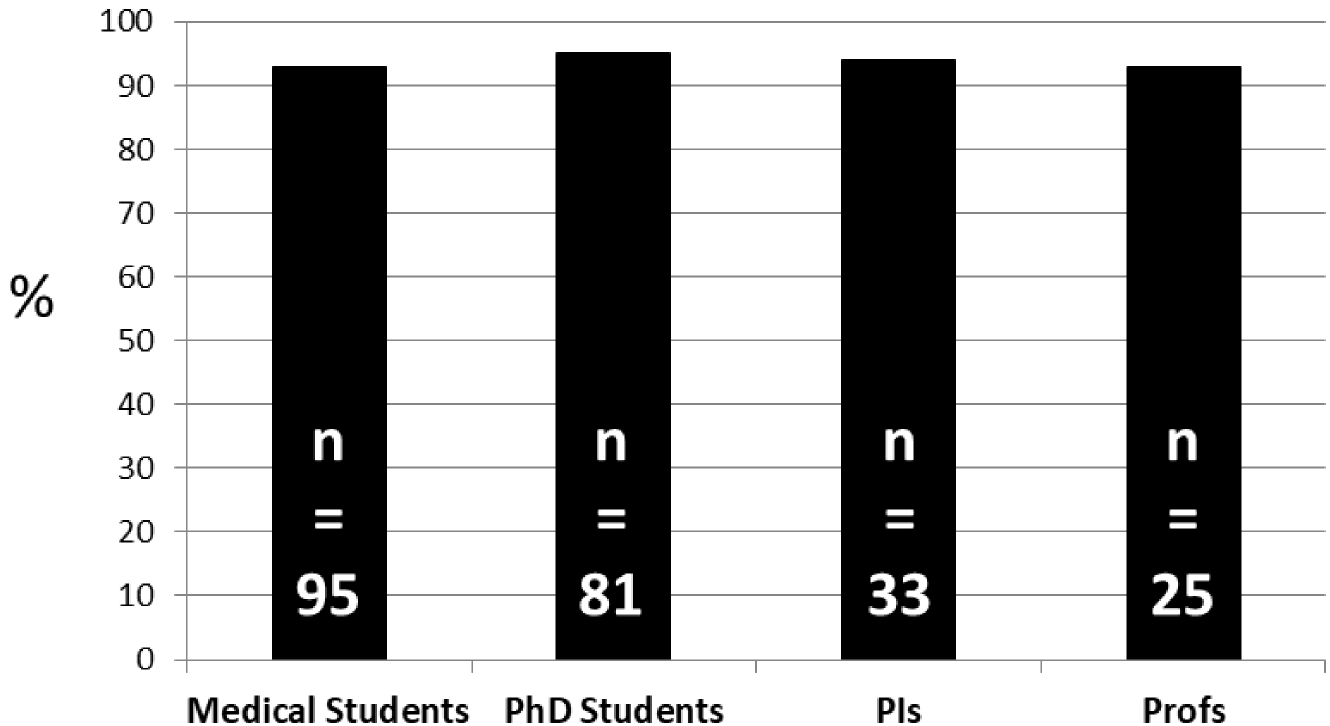
BIH QUEST
Center for Responsible Research

Statistical illiteracy and misconceptions: *‚Statistical significance‘*

Q (free text): What does $p<0.05$ actually mean?

The probability that my result is a fluke (my hypothesis was wrong, the drug doesn't work, etc.), is below 5 %...‘



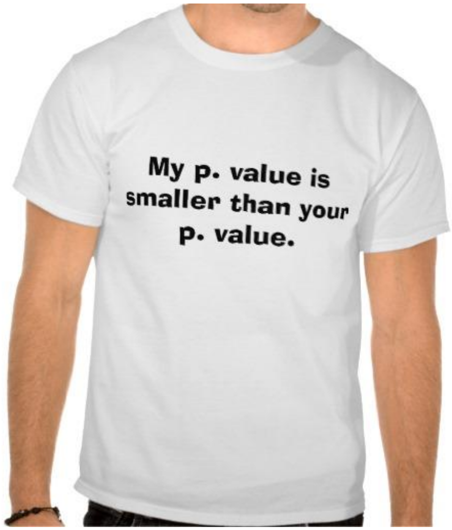


Source: Survey among participants of my seminar:
What you always wanted to know about the p-value, but did’nt dare to ask

BIH QUEST
Center for Responsible Research

Statistical illiteracy and misconceptions, e.g. regarding *p*:

- Belief that the p-value is negative (positive) predictive value
- The chance of replication exceeds 95%
- The chance that the result is a false positive is 5%
- There is a 95% chance that the alternative hypothesis is true (there is a high probability that the effect is real)
- The probability that the null hypothesis is true (that is, the probability of 'no effect' is 5%)
- etc.



Webinar: Dirnagl U. https://elifesciences.org/inside-elife/8f57d24c/what-you-always-wanted-to-know-about-the-p-value-but-were-afraid-to-ask

BIH QUEST
Center for Responsible Research

From the official exam questions for medical doctors:

Zunächst ist zu klären, ob bzgl. des Therapieansprechens tatsächlich ein Unterschied zwischen den Behandlungsgruppen vorliegt.

FEEDBACK GEBEN

A	Verblindung der Patienten gegenüber der Studienmedikation	15%	–
B	Methode der verdeckten Randomisierung (Concealment of allocation)	11%	–
C	Methode der Intention-to-Treat-Analyse	19%	–
D	Methode des Follow-up der Patienten	5%	–
E	statistische Signifikanz der Ergebnisunterschiede	51%	✓

Richtig! Wenn ein Ergebnisunterschied in der Stichprobe statistisch signifikant ist, beruht er mit hoher Wahrscheinlichkeit (i.d.R. ≥95%) auf einem „echten" Unterschied und nicht bloß auf Zufall. Bei der Beurteilung der statistischen Signifikanz hilft die Betrachtung des zugehörigen p-Wertes: Je niedriger der p-Wert ist, desto eher liegt ein „echter" Unterschied vor, der auf die Grundgesamtheit übertragen werden kann.

ZUSÄTZLICHE INFORMATIONEN

Angewandte Statistik

FEEDBACK GEBEN

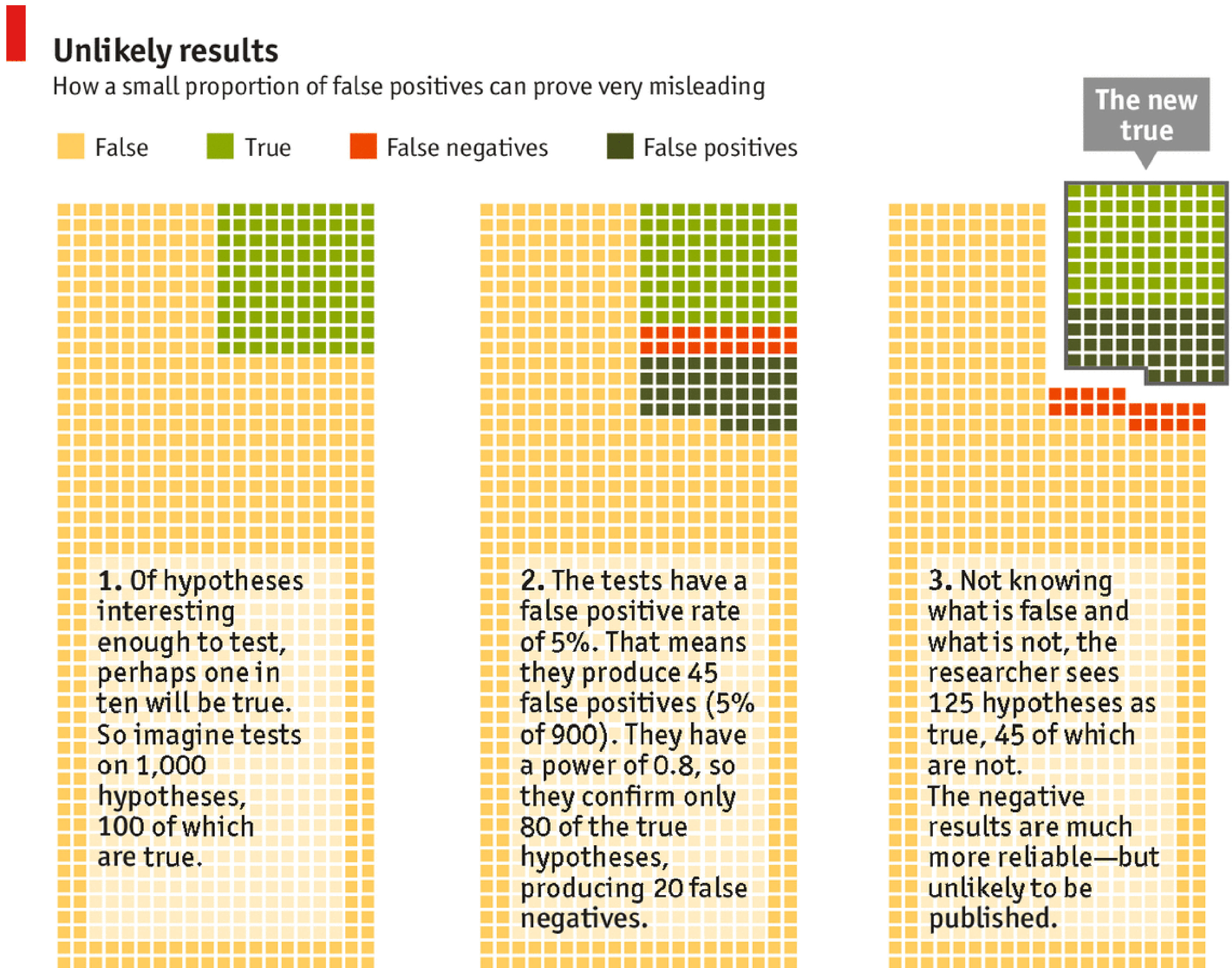
PRÜFUNGSMODUS AKTIVIEREN

FRAGE ZURÜCKSETZEN

"When a difference in results within the sample is statistically significant, it is highly likely (usually > 95%) to be due to a real difference and not merely due to chance. The lower the p-value, the more likely there is a real difference that can be generalized to the population."

Source https://www.amboss.com/de/), from IMPP)

BIH QUEST
Center for Responsible Research



Economist. Animated version: https://www.facebook.com/TheEconomist/videos/unlikely-results/10154245084204060/

P-values and claiming discovery

- RA Fisher: 5% (1:20) = ‚worth a look‘
- p-value is not a positive predictive value
- Inverse relationship of prior probability (base rate) of hypothesis and false positive rate

Statistical thresholds for claiming discoveries too low (The tale of a hog cycle…)

despite the awesome pre-eminence this method has attained in our journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428. DOI: 10.1037/h0042040

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.

https://www.nature.com/articles/s41562-017-0189-z

nature
human behaviour

Correspondence | Published: 25 September 2017

Remove, rather than redefine, statistical significance

Valentin Amrhein & Sander Greenland

Nature Human Behaviour **2**, 4 (2018) | Download Citation

https://www.nature.com/articles/s41562-017-0224-0

European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-019-04467-5

EDITORIAL

The *p* value wars (again)

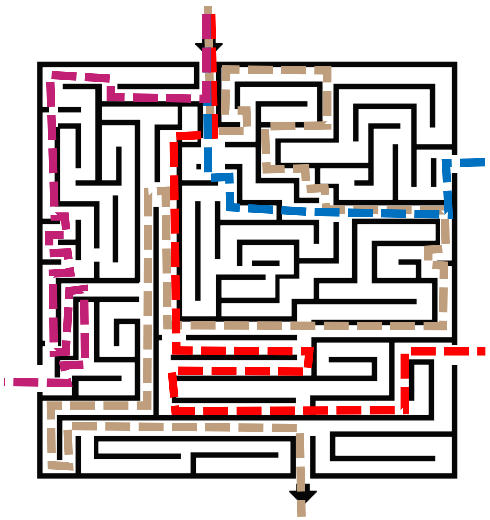
Ulrich Dirnagl[1,2]

https://link.springer.com/article/10.1007%2Fs00259-019-04467-5

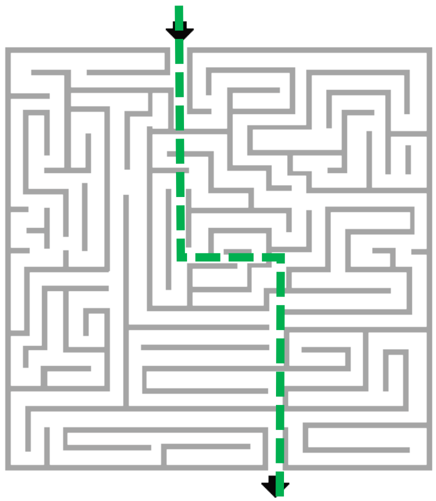
BIH QUEST
Center for Responsible Research

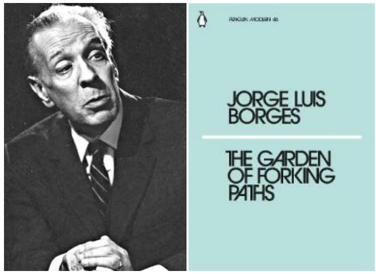
The labyrinth of the 'garden of forking paths'…

What happened in the project
(or could have happened…)



How the project was ‚sold'
in the publication
(‚Next, we…'- narrative)





Andrew Gelman† and Eric Loken‡

http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Blog post: http://bit.ly/2JzblTR

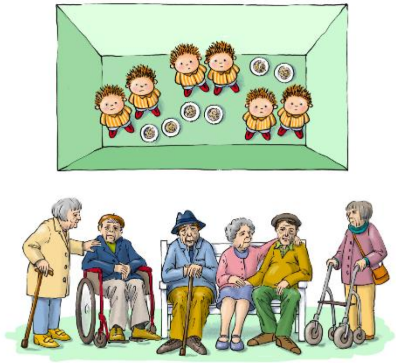
BIH QUEST
Center for Responsible Research

The perfect storm

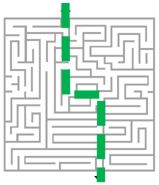
Biological complexity

Low internal validity

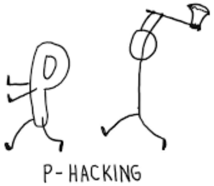
Publication bias



Low external validity



Cherry picking,
Story telling



Statistical blunders

Small n‘s, low power

Remedies

https://reproducibilitea.org/

Distinguishing between exploration and confirmation

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Perspective

Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation

Jonathan Kimmelman[1]*, Jeffrey S. Mogil[2], Ulrich Dirnagl[3,4,5]

PLoS Biol. (2014) 12:e1001863.

https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001863

BIH QUEST
Center for Responsible Research

Exploration/Discovery vs. Confirmatory (knowledge claiming) research

Exploration: Generates hypotheses and does <u>not lead to a formal knowledge claim.</u>

Hypothesis testing / Confirmatory / Knowledge claiming experiment: A clear, predefined hypothesis, including a clear predefined primary outcome measure to test the hypothesis and a predefined and appropriate statistical test. The proposed sample size should be stated, along with a justification based on the statistical power to detect a biologically important effect.

A given study can involve hypothesis-testing <u>and</u> exploratory parts, for instance by defining one primary endpoint (hypothesis-testing), with all other measured endpoints being exploratory

There is a one-way street between confirmatory and exploratory experiments: if you find interesting results which contradict your hypothesis, a confirmatory experiment can turn into an exploratory experiment. However, an exploratory experiment can never become confirmatory.

	Exploratory	Confirmatory
Hypothesis	(+)	+++
Establish pathophysiology („knowledge claim“)	+++	(+)
Sequence and details of experiments established at onset	(+)	+++
Primary endpoint	-	++
Sample size calculation	(+)	+++
Blinding	+++	+++
Randomization	+++	+++
External validity (aging, comorbidities, etc.)	-	++
In/Exclusion criteria	++	+++
Test statistics	+	+++
Preregistration	(-)	+++
Sensitivity (Type II error) Find what might work	++	+
Specificity (Type I error) Weed out false positives	+	+++

Dirnagl, U. https://www.ahajournals.org/doi/full/10.1161/STROKEAHA.116.013244

BIH QUEST
Center for Responsible Research

frontiers in Neurology | Stroke

Impact Factor 4.
More on impact

SECTION ABOUT ARTICLES RESEARCH TOPICS FOR AUTHORS EDITORIAL BOARD ARTICLE ALERT

‹ Articles

ORIGINAL RESEARCH article

Front. Neurol., 19 November 2018 | https://doi.org/10.3389/fneur.2018.00937

Check for updates

Exploratory Investigation of Intestinal Function and Bacterial Translocation After Focal Cerebral Ischemia in the Mouse

Naoki Oyama[1,2†], Katarzyna Winek[1,2,3*†‡], P
Claudia Dames[4], Martina Werich[5], Olivia Ker
Meisel[1,2,3,7‡] and Ulrich Dirnagl[1,2,3,7,8,9‡]

Original Article

JCBFM

Journal of Cerebral Blood Flow & Metabolism
2020, Vol. 40(2) 276–287

DOI: 10.1177/0271678X19827251
journals.sagepub.com/home/jcbfm
SAGE

An exploratory investigation of brain collateral circulation plasticity after cerebral ischemia in two experimental C57BL/6 mouse models

Marco Foddis[1,*], Katarzyna Winek[1,*], Kajetan Bentele[2], Susanne Mueller[1,3], Sonja Blumenau[1], Nadine Reichhart N[4], Sergio Crespo-Garcia[4], Dermot Harnett[2], Andranik Ivanov[2], Andreas Meisel[1], Antonia Joussen[4], Olaf Strauss[4], Dieter Beule[2], Ulrich Dirnagl[1,5] and Celeste Sassi[1]

eLife

FEATURE ARTICLE

SCIENCE FORUM

Improving preclinical studies through replications

Abstract The purpose of preclinical research is to inform the development of novel diagnostics or therapeutics, and the results of experiments on animal models of disease often inform the decision to conduct studies in humans. However, a substantial number of clinical trials fail, even when preclinical studies have apparently demonstrated the efficacy of a given intervention. A number of large-scale replication studies are currently trying to identify the factors that influence the robustness of preclinical research. Here, we discuss replications in the context of preclinical research trajectories, and argue that increasing validity should be a priority when selecting experiments to replicate and when performing the replication. We conclude that systematically improving three domains of validity – internal, external and translational – will result in a more efficient allocation of resources, will be more ethical, and will ultimately increase the chances of successful translation.

NATASCHA INGRID DRUDE†, LORENA MARTINEZ GAMBOA†, MEGGIE DANZIGER, ULRICH DIRNAGL AND ULF TOELCH*

https://elifesciences.org/articles/62101

BIH QUEST
Center for Responsible Research

More replication of results

Preconditioning with CpG-ODN1826 reduces ischemic brain injury in young male mice: a replication study

Kunjan R. Dave[1,2,3], Isabel Saul[1,2], Ami P. Raval[1,2,3], Miguel A. Perez-Pinzon[1,2,3]

[1]Peritz Scheinberg Cerebral Vascular Disease Research Laboratories, University of Miami School of Medicine, Miami, FL, USA.

[2]Department of Neurology, University of Miami School of Medicine, Miami, FL, USA.

[3]Neuroscience Program, University of Miami School of Medicine, Miami, FL, USA.

http://www.conditionmed.org/Data/View/6289

eLife

Replication Study: Intestinal inflammation targets cancer-inducing activity of the microbiota

Kathryn Eaton, Ali Pirani, Evan S Snitkin, Reproducibility Project: Cancer Biology*

Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, United States

https://elifesciences.org/articles/34364

Research Paper

PAIN

OPEN

Antihyperalgesic effects of Meteorin in the rat chronic constriction injury model: a replication study

Jennifer Y. Xie[a], Chaoling Qu[b], Gordon Munro[c], Kenneth A. Petersen[c], Frank Porreca[b,*]

doi: 10.1097/j.pain.0000000000001569

BIH QUEST
Center for Responsible Research

Team science: Preclinical randomized controlled multicenter trials

RESEARCH ARTICLE

STROKE

Results of a preclinical randomized controlled multicenter trial (pRCT): Anti-CD49d treatment for acute brain ischemia

Science Translational Medicine AAAS
2015;7:299ra121

https://www.science.org/doi/10.1126/scitranslmed.aaa9853

https://doi.org/10.1093/braincomms/fcad090 BRAIN COMMUNICATIONS 2023: Page 1 of 13 | 1
BRAIN COMMUNICATIONS

A preclinical randomized controlled multi-centre trial of anti-interleukin-17A treatment for acute ischaemic stroke

https://doi.org/10.1093/braincomms/fcad090

SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

STROKE

A multi-laboratory preclinical trial in rodents to assess treatment candidates for acute ischemic stroke

https://www.science.org/doi/full/10.1126/scitranslmed.adg8656

BIH QUEST
Center for Responsible Research

Preregistration of preclinical study protocols

- Limits unwarranted and/or undisclosed researcher‘s degrees of freedom‘
- Prevents ‚outcome switching‘
- Prevents HARKING
- Provides scooping protection
- Reduces publication bias
- Distinguishes between exploratory/discovery and knowldege claiming / confirmatory research
- …

(Pre) Registration of ‚exploratory‘ preclinical research?

PLOS BIOLOGY

PERSPECTIVE

Preregistration of exploratory research: Learning from the golden age of discovery

Ulrich Dirnagl *

QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Berlin, Germany

Citation: Dirnagl U (2020) Preregistration of exploratory research: Learning from the golden age of discovery. PLoS Biol 18(3): e3000690. https://doi.org/10.1371/journal.pbio.3000690

https://doi.org/10.1371/journal.pbio.3000690

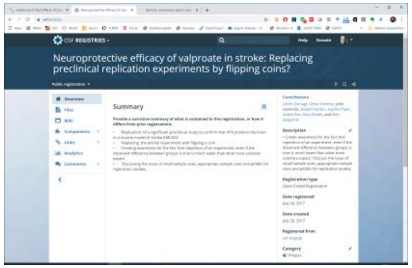
BIH QUEST
Center for Responsible Research

Preregistration of study protocols (preclinical)

All purpose registries
(**not** reviewed)

https://osf.io/

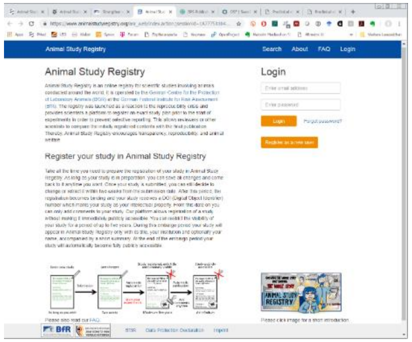
https://aspredicted.org/



Animal study registries (ASR)
(**not** reviewed)

German Centre for the Protection of Laboratory Animals
https://www.animalstudyregistry.org

Preclinicaltrials.eu https://preclinicaltrials.eu/



Timestamp servers / Blockchain
(**not** reviewed)

e.g. https://github.com/decred/dcrtimegui

Registered reports (Elife, PlosBiol, F1000Res etc.) (***reviewed!***)

BIH QUEST
Center for Responsible Research

Early statistical consultation

Ronald Fisher (1938)

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

Novel (more efficient) analytical approaches and study designs

nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-020-00792-3

Check for updates

Increasing the statistical power of animal experiments with historical control data

V. Bonapersona[1], H. Hoijtink[2], RELACS Consortium*, R. A. Sarabdjitsingh[1,13] and M. Joëls[1,3,13]

https://www.nature.com/articles/s41593-020-00792-3

https://youtu.be/vtWBQAlGrFI

BIH QUEST
Center for Responsible Research

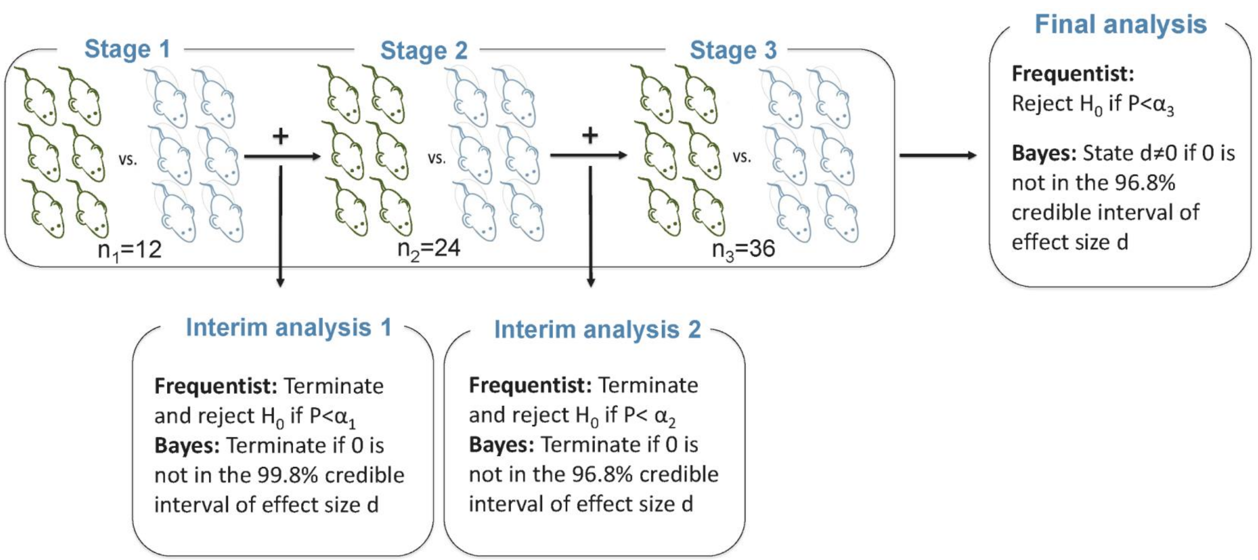
Novel (more efficient) analytical approaches and study designs

PLOS | BIOLOGY

PERSPECTIVE

Increasing efficiency of preclinical research by group sequential designs

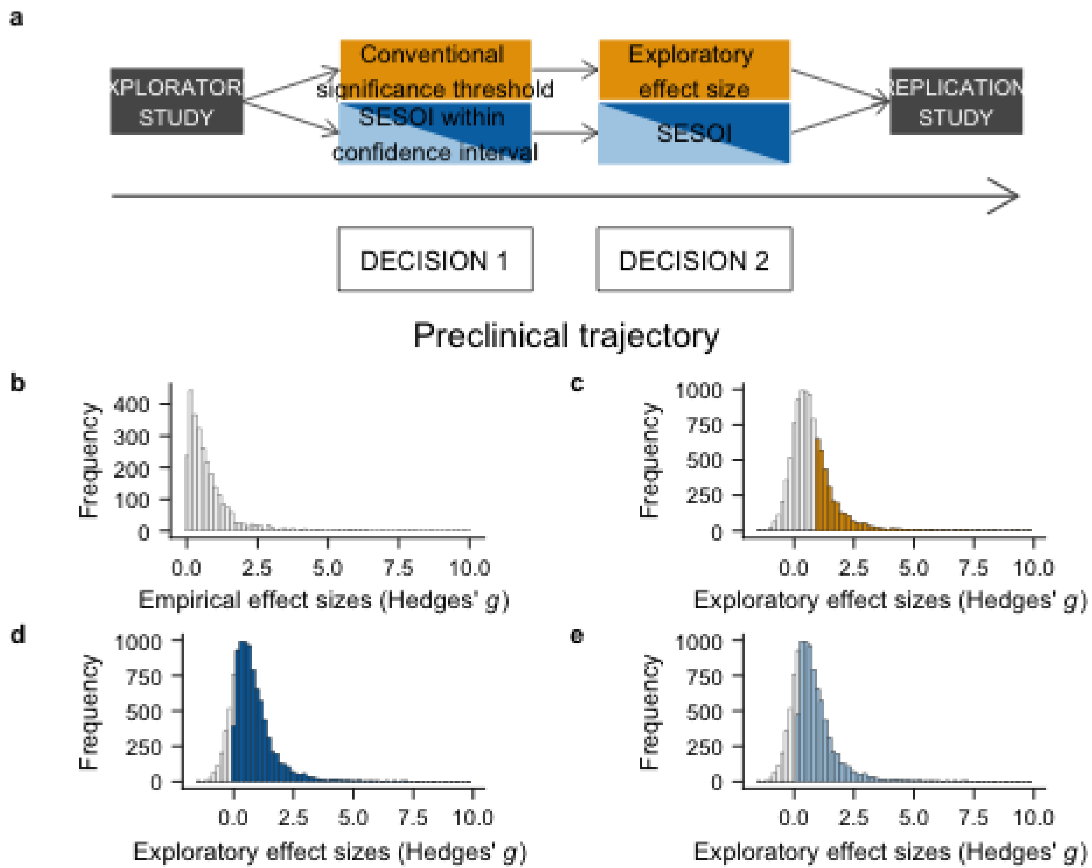
Konrad Neumann[1☯], Ulrike Grittner[1,2☯*], Sophie K. Piper[1,2,3], Andre Rex[2,4], Oscar Florez-Vargas[5], George Karystianis[6], Alice Schneider[1,2], Ian Wellwood[2,7], Bob Siegerink[2,8], John P. A. Ioannidis[9], Jonathan Kimmelman[10], Ulrich Dirnagl[2,3,4,8,11,12]



https://doi.org/10.1371/journal.pbio.2001307

BIH QUEST
Center for Responsible Research

Increasing discovery rates in preclinical research through optimised statistical decision criteria (smallest effect size of interest - SESOI)



Balancing sensitivity and specificity in preclinical research

Meggie Danziger, Anja Collazo, Ulrich Dirnagl, Ulf Toelch

https://doi.org/10.1101/2022.01.17.476585

BIH QUEST
Center for Responsible Research

Causal Inference / DAGs in preclinical research

Journal of Cerebral Blood Flow & Metabolism
OnlineFirst

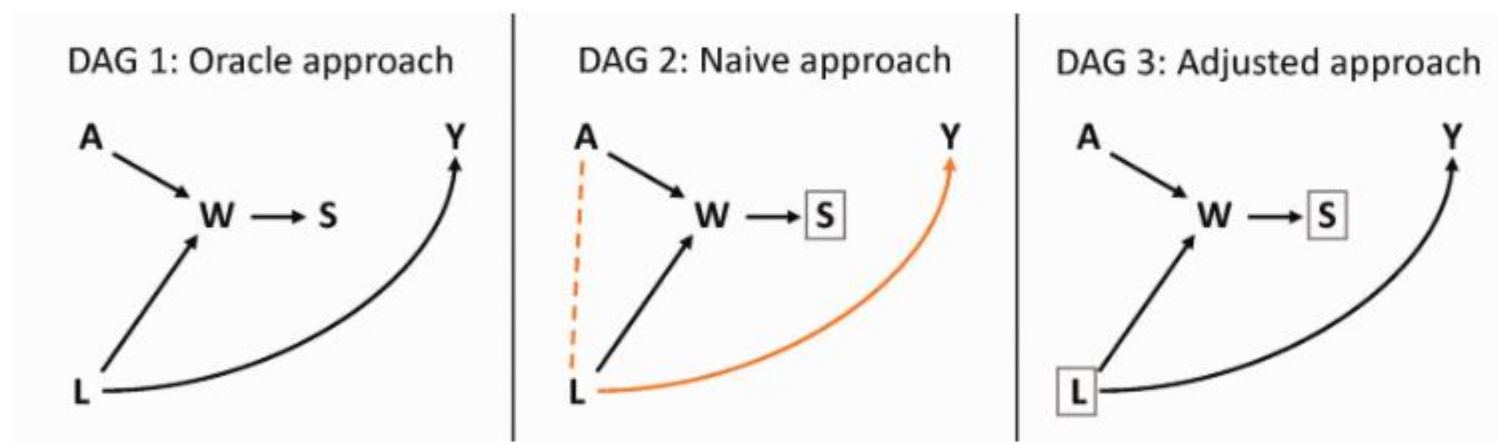
https://doi.org/10.1177/0271678X241275760

Journal of Cerebral Blood Flow & Metabolism

Original Article

Rethinking animal attrition in preclinical research: Expressing causal mechanisms of selection bias using directed acyclic graphs

Anja Collazo [1,2], Hans-Georg Kuhn [2,3], Tobias Kurth [2], Marco Piccininni [2,4], and Jessica L Rohmann [2,4,5]



https://doi.org/10.1177/0271678X241275

COLLABORATIVE VISIONS BIOSTATISTICS SYMPOSIUM

STRAUSBERG, GERMANY
29^{TH} SEPT – 1^{ST} OCT 2024

The Lakeside Burghotel zu Strausberg, Gielsdorfer Chaussee 6, 15344 Strausberg.

Steven Goodman, David Allison, Shai Silberberg, Natasha Karp, Robert Nadon ….

BIH QUEST
Center for Responsible Research

Presymposium survey

Institutions

- Insufficient training in experimental design and data analysis
- Insufficient support by biostatisticians (capacity)
- Faculty evaluation does not include rigor of research (including proper use of statistics)
- Teaching purely ‘technical’ (How to do an ANOVA or regression…), but not about concepts
- Experiments first, apply statistics post hoc “ we will sort the statistics later”

Scientists

- Lack of competence or support
- Being a part of a research culture that incentivises a focus on outputs over process
- Poor communication between wet-lab scientists and statisticians leads to misunderstandings
- Projects start without involvement of a biostatistician.

Funders

- Study design/stats underrepresented in proposals
- Lack of stats competence by referees
- Grant evaluation processes that do not give enough weight during assessment to methodological rigour
- Lack of career path/jobs for non-clinical statisticians.

Publishers

- Not enough focus by journals and their editorial and peer review processes on methods and methodological rigour, and too much focus/reward for ‘positive’ results
- Lack of competent reviewers
- Switching of analysis, cherry picking, no preregistered study/analysis plans

Other relevant stakeholders include: Learned societies, Policy makers, Open science/Reproducibility Initiatives, Investors, Regulatory Authorities

The root cause: An academic incentive structure, which prioritizes publishing eye-catching results in high-impact journals—at the expense of scientific rigor and robustness.

The academic reputation economy

Publishers provide the 'currency' in the form of a hierarchy of journals ('different exchange rates') - this brings them enormous profits

Academia receives a practical, pseudo-objective criterion for 'performance'. Research is now conducted not only for the sake of gaining knowledge, but for a paper in Nature

This stabilizes the system (even universities and entire countries now use this currency in rankings). It thus becomes a gigantic hurdle for those who want to change it.

BIH QUEST
Center for Responsible Research

Conclusions

- Statistical misconceptions, flawed experimental design, and undisclosed or unwarranted researchers' degrees of freedom are key contributors to high attrition rates and lack of reproducibility in preclinical research.
- The issue is not merely a matter of insufficient education, professional support, or resources; it is fundamentally a cultural problem within the biomedical scientific community.
- Increased education, better support, and innovative statistical methods can mitigate these issues but will not fully resolve the underlying problem.
- The root cause lies in the existing academic incentive structure, which prioritizes publishing eye-catching results in high-impact journals—often at the expense of scientific rigor and robustness.

Slide download http://bit.ly/dirnaglncs