Biomedical research: Is failed replication always a bad thing?

Malcolm Macleod

Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies

and

University of Edinburgh

Disclosures

- UK Commission for Human Medicines
- EMA Neurology SAG

- UK Animals in Science Committee

- Independent Statistical Standing Committee, CHDI Foundation
- Project co-ordinator, EQIPD IMI

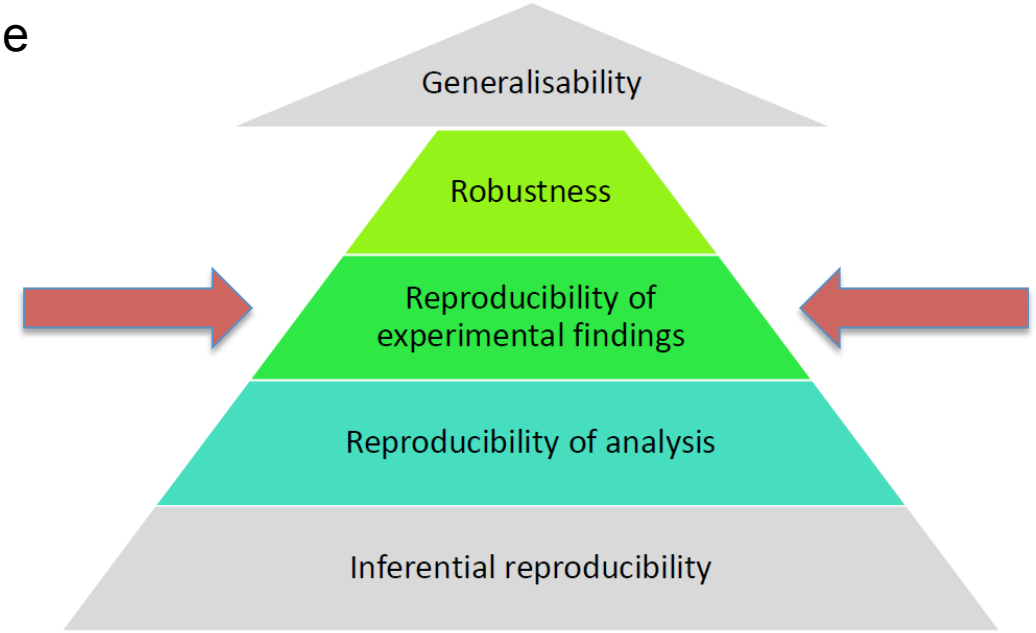
What is research?

- Analysis and interpretation of observations
- Setting may be spontaneous or experimental
- Leads to a knowledge claim
- Usually involves a statistical analysis

Reproducibility and replication

"Reproducibility" related to the re-analysis of existing data following the same analytical procedures.

"Replication" was held to require the collection of new data, following the same methods.



Replication studies

1. Retrospective – Pharmaceutical companies sharing their historical experience when they have attempted replication
   - – Bayer 33% of 67
   - – Amgen 11% of 53

   Selection bias (2 companies out of ?)

   ? Recall Bias

Replication studies

2. Prospective - Academic led, great attention given to faithfulness to original study design, adequate statistical power, preregistration

– Psychology	36% of 97	ES_R=49%
– Cancer biology	40% of 10	
– Economics	61% of 18	ES_R=66%
– Social sciences	62% of 21	ES_R=54%

? Selection bias (how did they choose what to try to replicate?)

Claim

- Lack of reproducibility of experimental findings has been observed across such a wide variety of settings that it can be considered a general phenomenon
- Therefore, unless a field can demonstrate that it doesn't have a problem, it is reasonable to expect that it does

ORIGINAL Finding	REPLICATION Finding
INCORRECT	CORRECT
CORRECT	INCORRECT
CORRECT	CORRECT

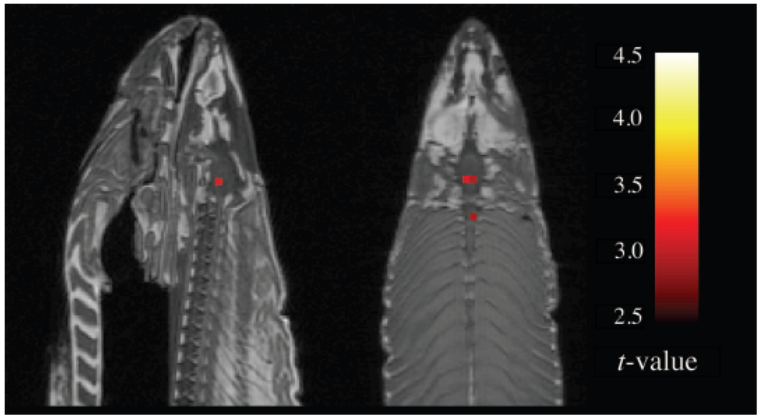
I am not in the office at the moment. Send any work to be translated.

Journal of Serendipitous and Unexpected Results

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1]*, Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.



The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing.

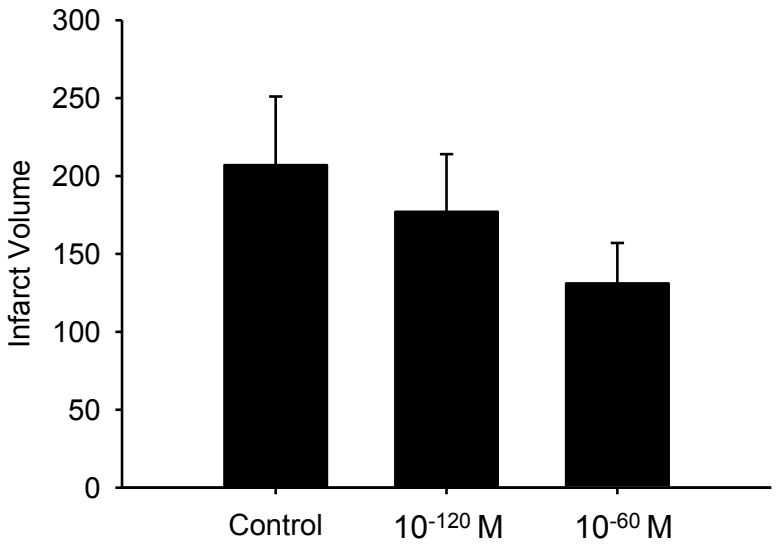
Several active voxels were observed in a cluster located within the salmon's brain cavity (see Fig. 1). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$.

Either we have stumbled onto a rather amazing discovery in terms of post-mortem ichthyological cognition, or there is something a bit off with regard to our uncorrected statistical approach.

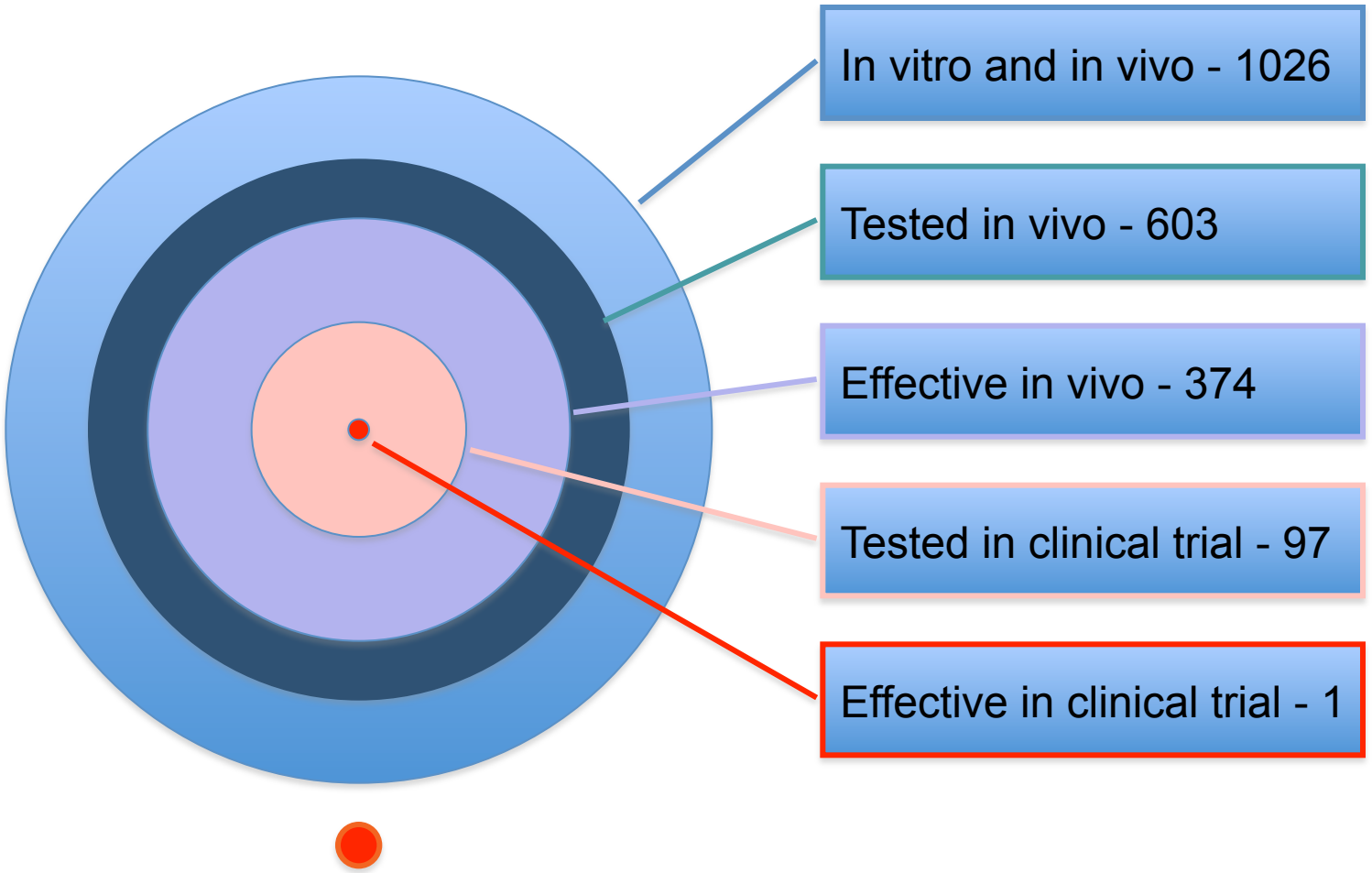
Winner of the 2012 Ignoble Prize for Neuroscience

Treatment of experimental stroke with low-dose glutamate and homeopathic Arnica montana*

W. Jonas[1], Y. Lin[2], A. Williams[2], F. Tortella[2], R. Tuma[3]
[1] Uniformed Services University of the Health Sciences, Bethesda, Maryland
[2] Walter Reed Army Institute of Research, Washington, D.C.
[3] Temple University, Philadelphia, PA



1026 interventions in experimental stroke



O' Collins et al, 2006

The originator study is incorrect

1. Most published research is false

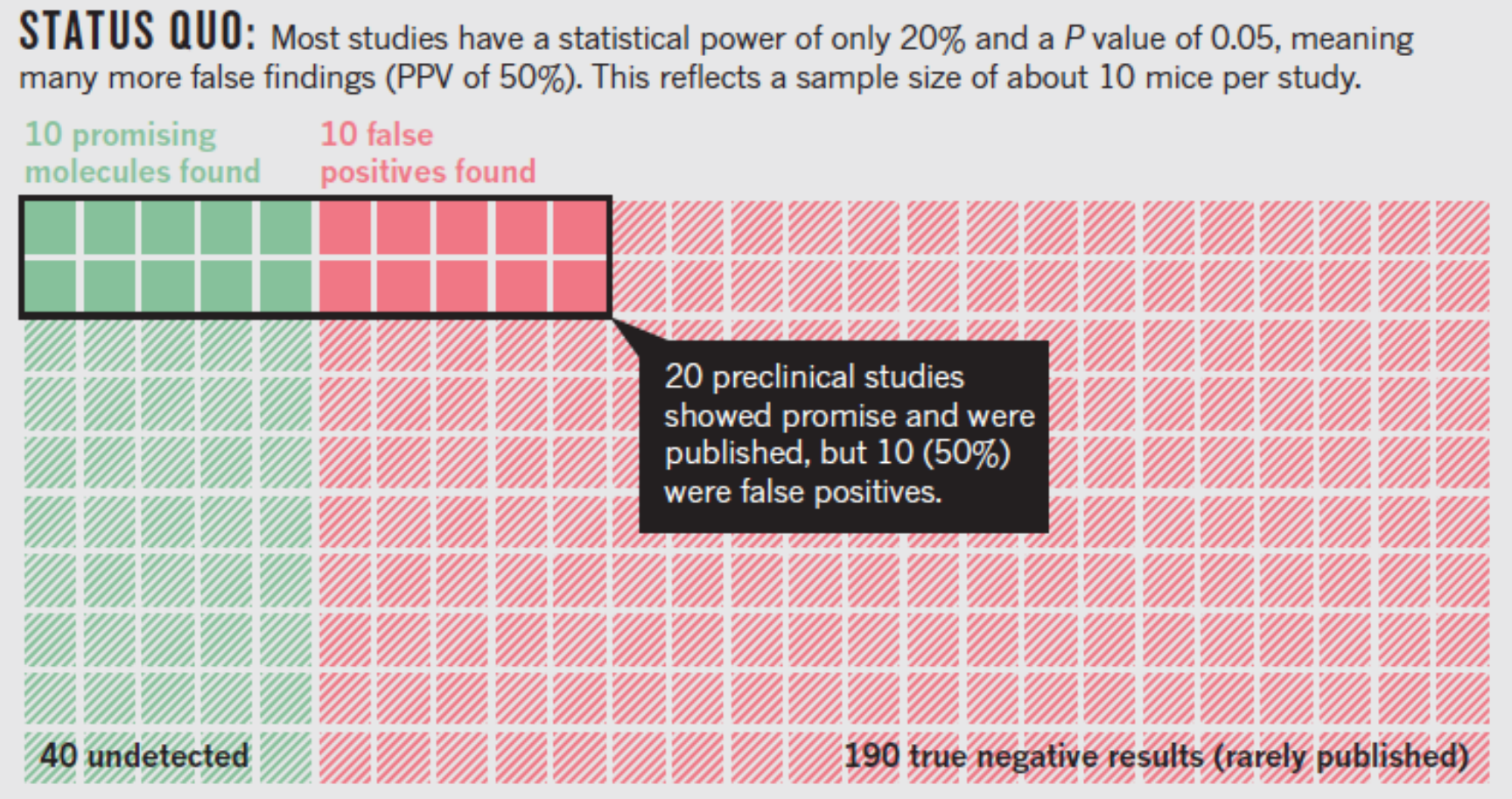
Power = 20%

Alpha = 5%

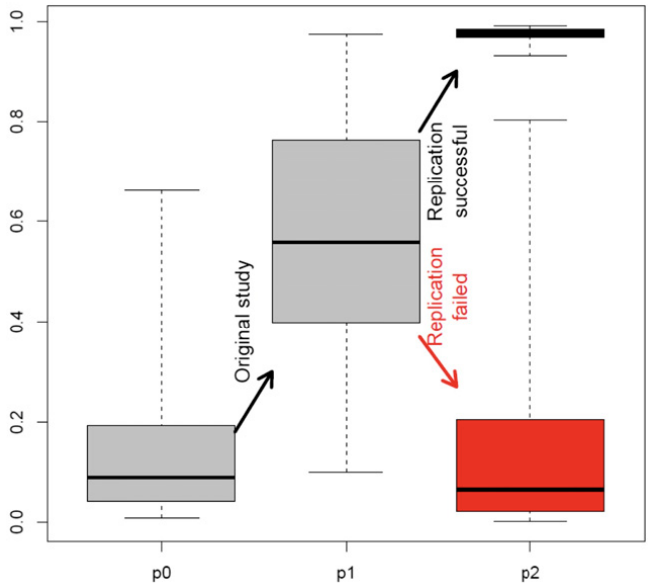
Proportion of studies that are truly positive ("prior") = 10%

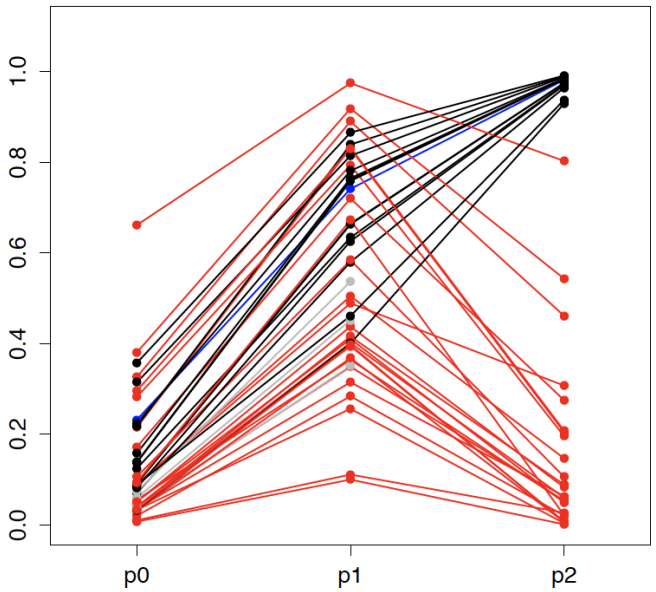
Chance that "significant" finding is true = 31%

Take 250 in vivo studies …



Psychology Replication Project (hat tip Anna Dreber)





For each study,

- p1 is the ”prior” for the replication effort (derived from market)
- p0 is the **calculated** original ”prior”
- p2 is the posterior

...

- p1 ∝
  - strength of original evidence
  - expert critical appraisal
- For each study, also know power of replication study, so can predict probability of successful replication (=p1*power)
- Averaging across 41 studies,

  p(rep) = 0.53, p(non-rep) = 0.47

  ∴ expected non-replication = 19 studies

  observed non-replication = 25 studies

  attributable non-replication = 76%

The originator study is incorrect

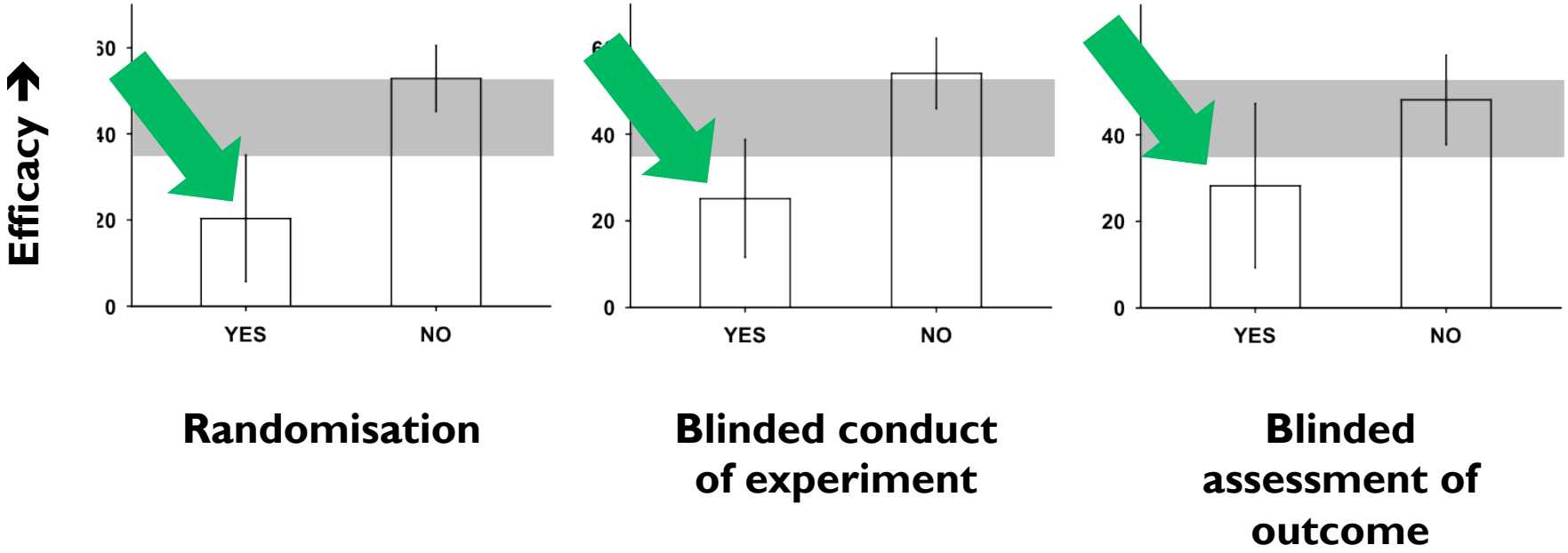
1. Most published research is false
2. HARKing
3. Flexibility in data analysis (p-hacking)
4. Publication bias, selective outcome reporting bias

The originator study reports inflated effect size estimates

1. Designs which introduce risks of bias
2. Publication bias, selective outcome reporting bias

Risk of bias in animal studies

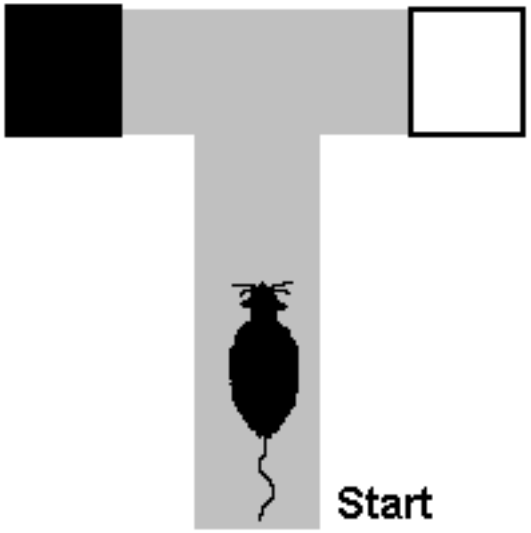
- Infarct Volume
  - 11 publications, 29 experiments, 408 animals
  - Improved outcome by 44% (35-53%)



Macleod et al, 2008

You can usually find what you're looking for …

- 12 graduate psychology students
- 5 day experiment: rats in T maze with dark arm alternating at random, and the dark arm always reinforced
- 2 groups – "Maze Bright" and "Maze dull"

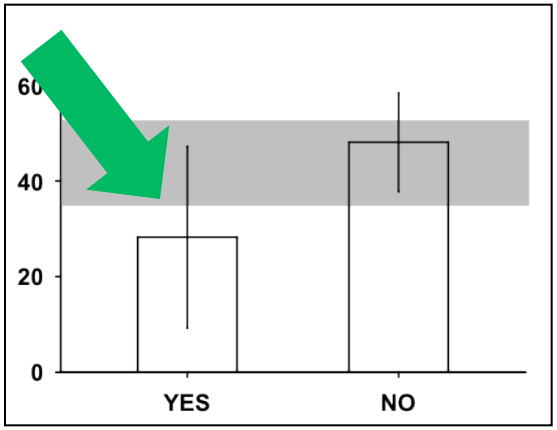


Group	Day 1	Day 2	Day 3	Day 4	Day 5
"Maze bright"	1.33	1.60	2.60	2.83	3.26
"Maze dull"	0.72	1.10	2.23	1.83	1.83
Δ	+0.60	+0.50	+0.37	+1.00	+1.43

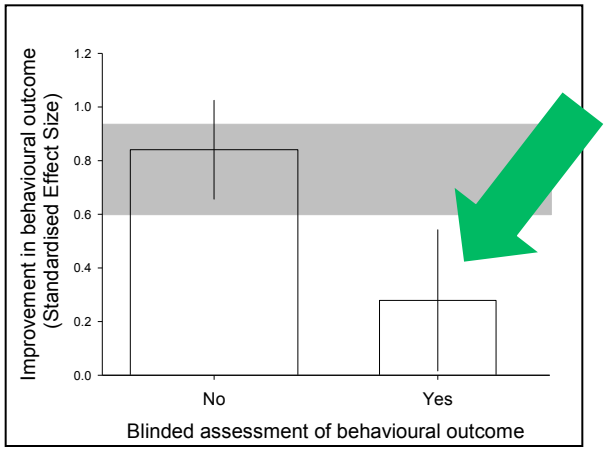
Rosenthal and Fode (1963), Behav Sci 8, 183-9

Evidence from various neuroscience domains …

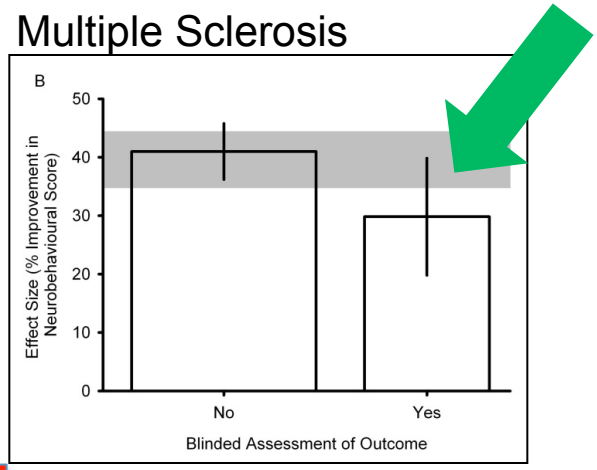
Stroke



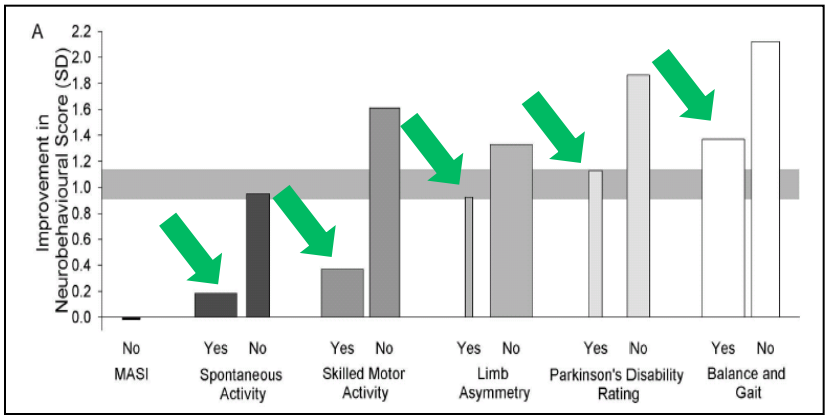
Alzheimer´s disease



Multiple Sclerosis



Parkinson´s disease



The scale of the problem

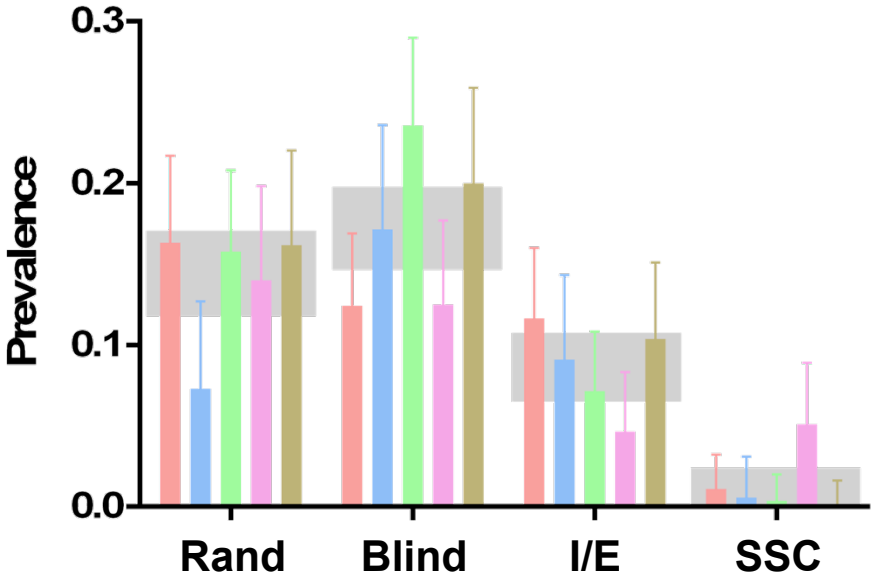
RAE 1173

rae2008
Research Assessment Exercise

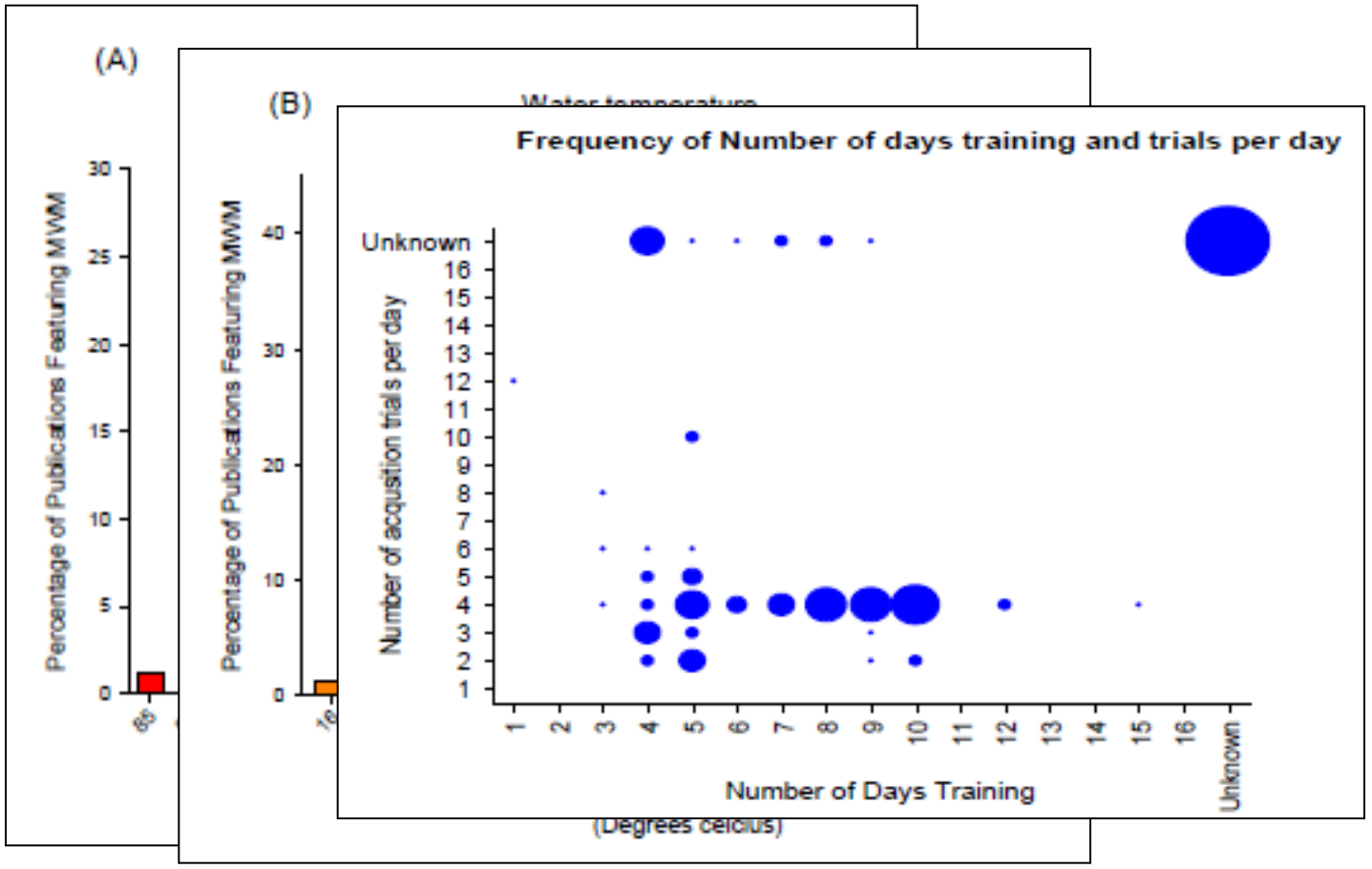
"an outstanding contribution to the internationally excellent position of the UK in biomedical science and clinical/ translational research."

"impressed by the strength within the basic neurosciences that were returned …particular in the areas of behavioural, cellular and molecular neuroscience"

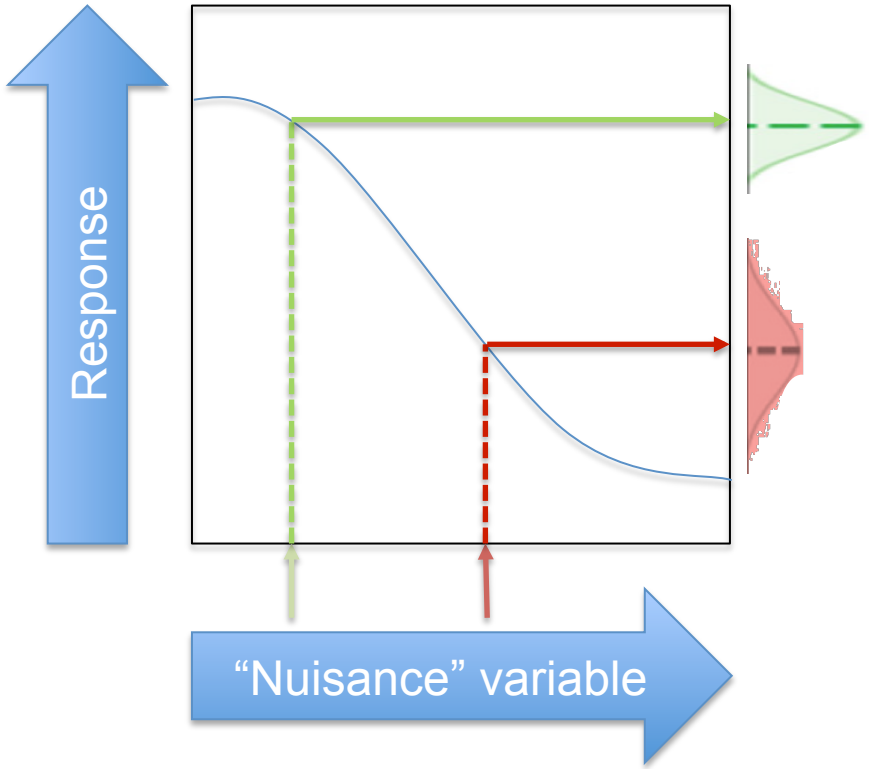
1173 publications using non human animals, published in 2009 or 2010, from 5 leading UK universities



Impossible replication – the Morris Water Maze



Both studies may be correct
Reaction norms (Voelkl 2016)



Lifespan in worms

Source of variation	Developmental Rate	Fertility
Genetic	83.1%	63.3%
Between labs	8.3%	7.9%
Within labs	3.8%	5.6%
Individual	4.8%	23.3%

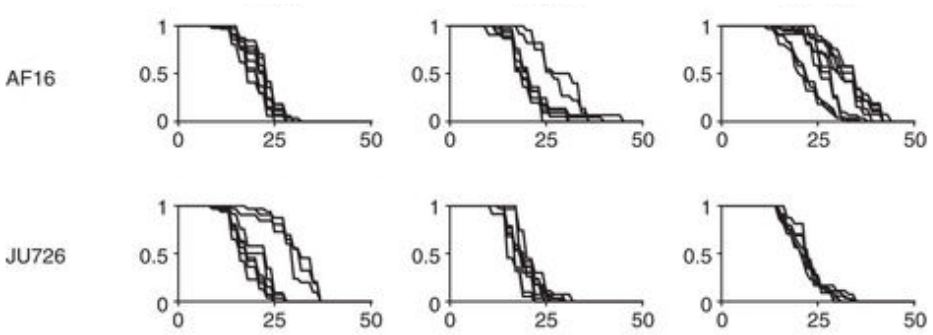
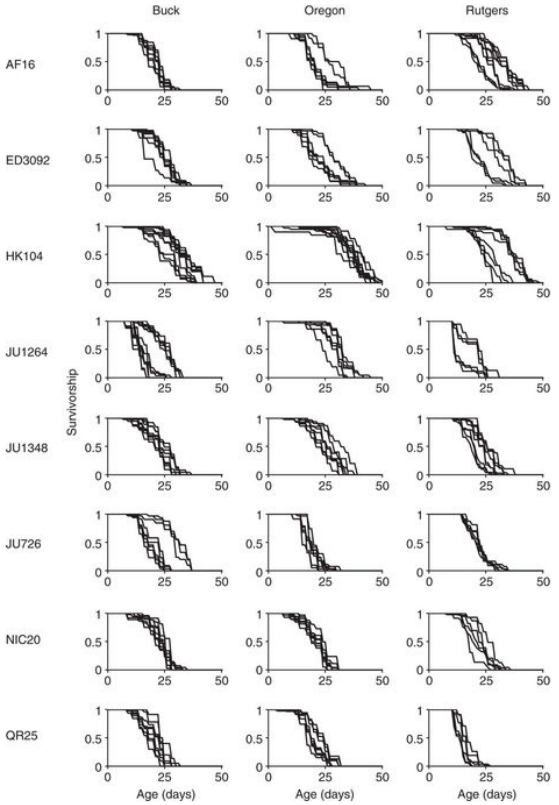


Figure 3: Variation in longevity within labs for each replicate plate for eight natural isolates of *C. briggsae*.

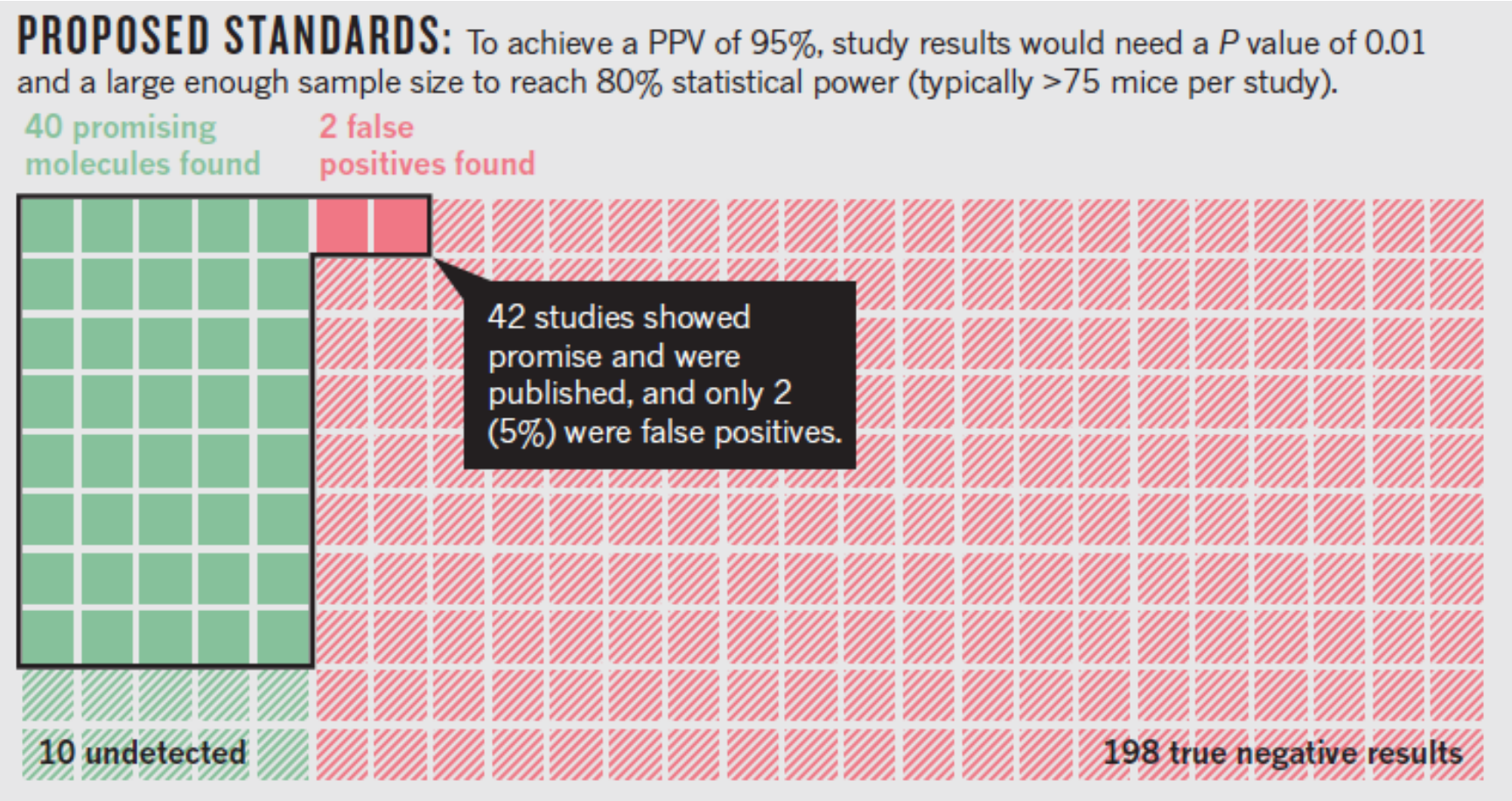


Lucanic et al Nature Comms 2017

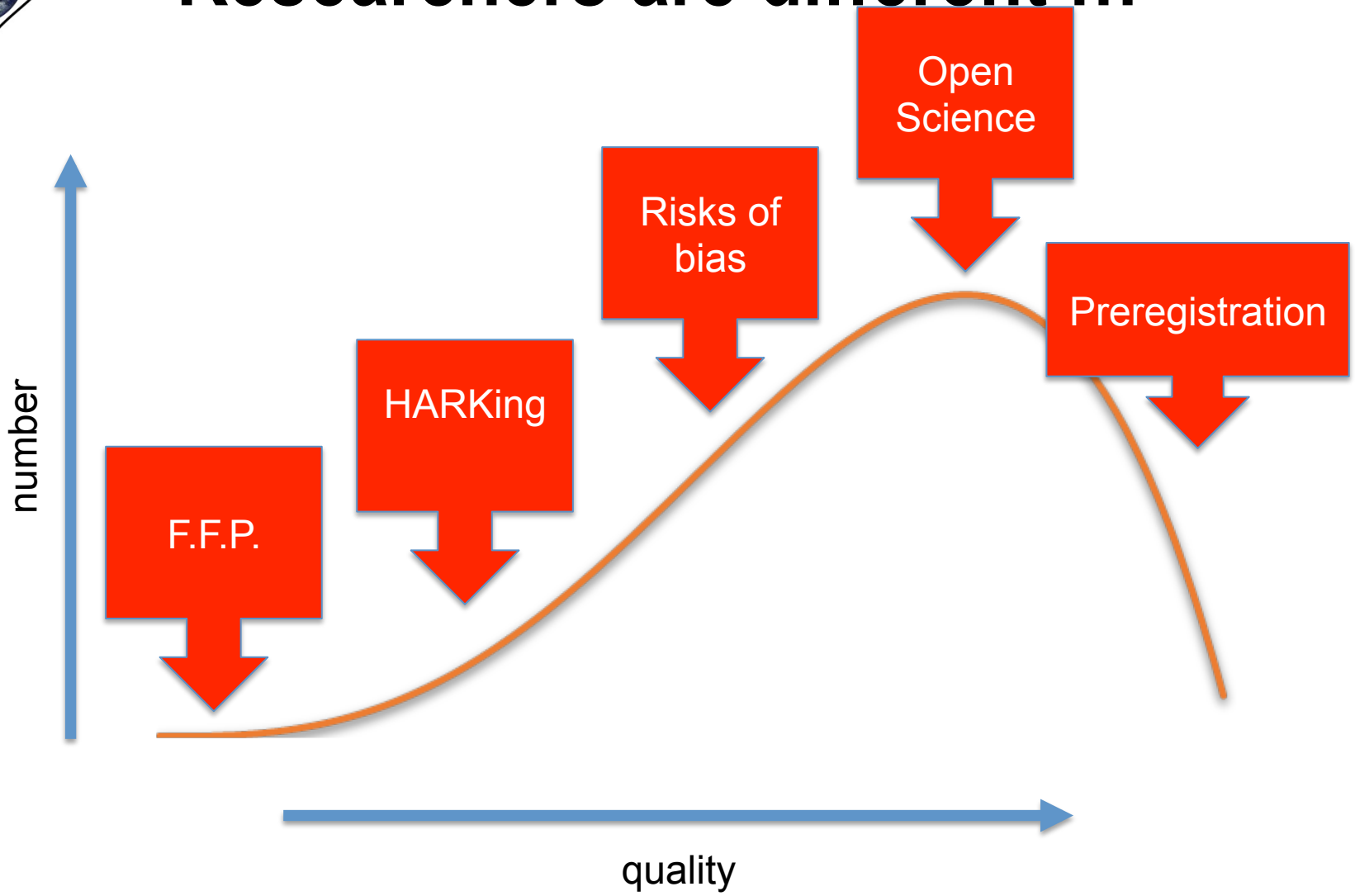
What should we do?

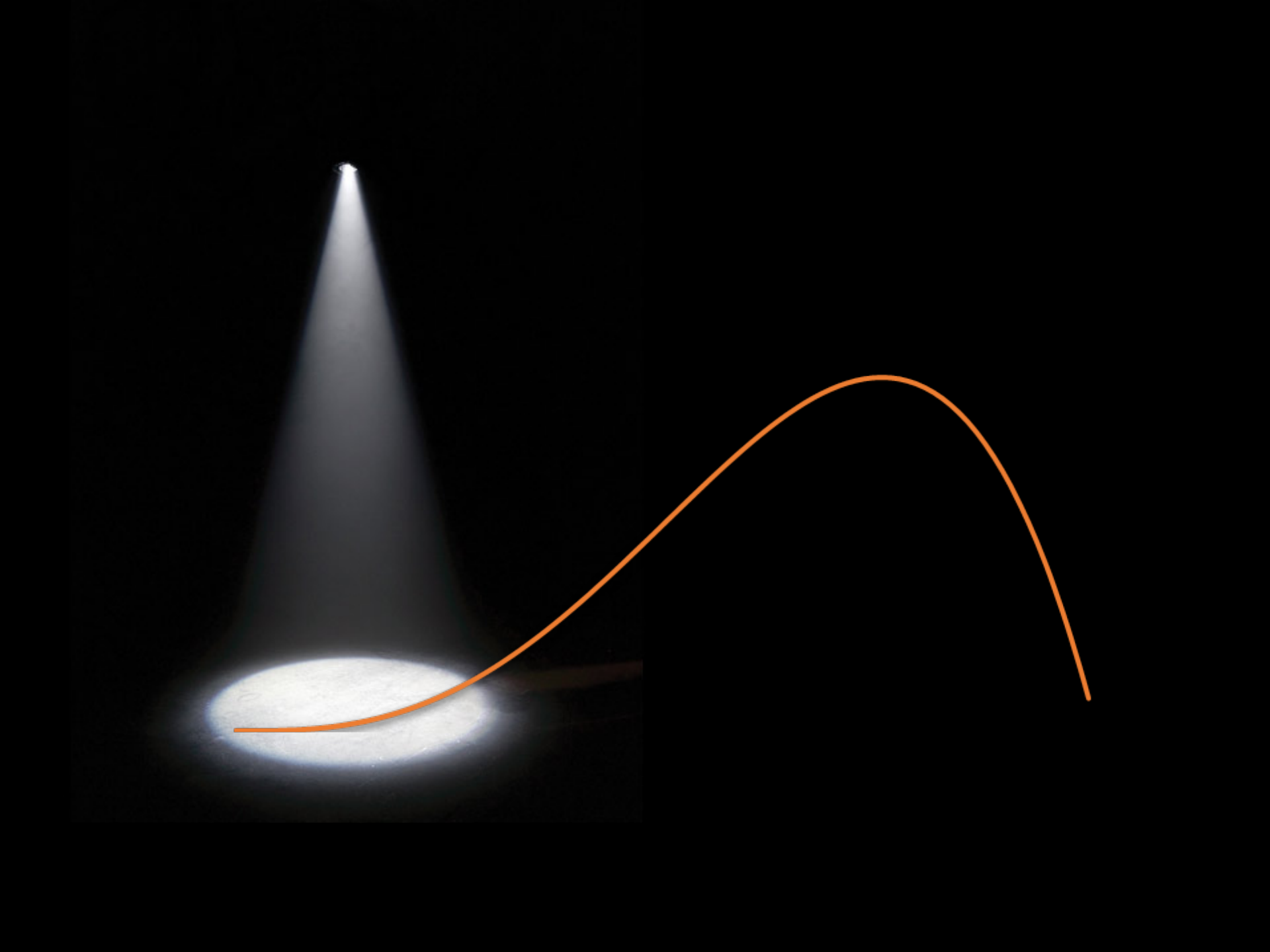
1. increase the probability that published research is true
2. establish a framework to select efficiently which research findings we should attempt to replicate
3. develop strategies to evaluate the robustness of key research findings

…with p<0.01, power @ 80%

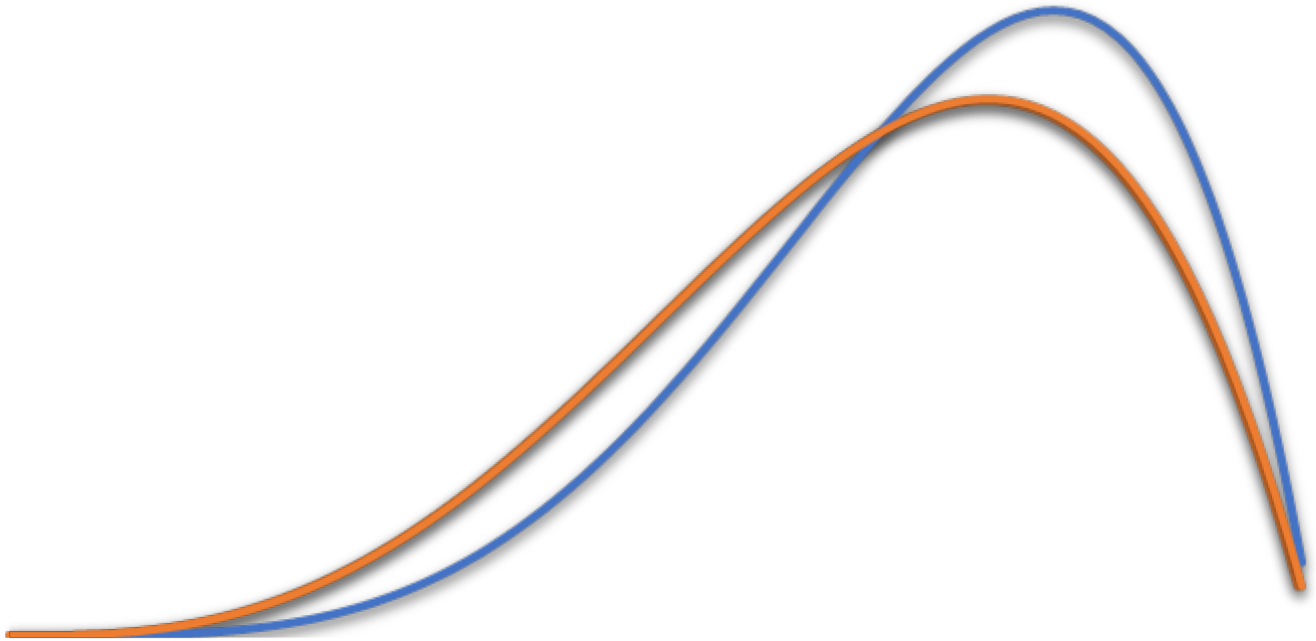


Researchers are different …





Research Improvement Strategy



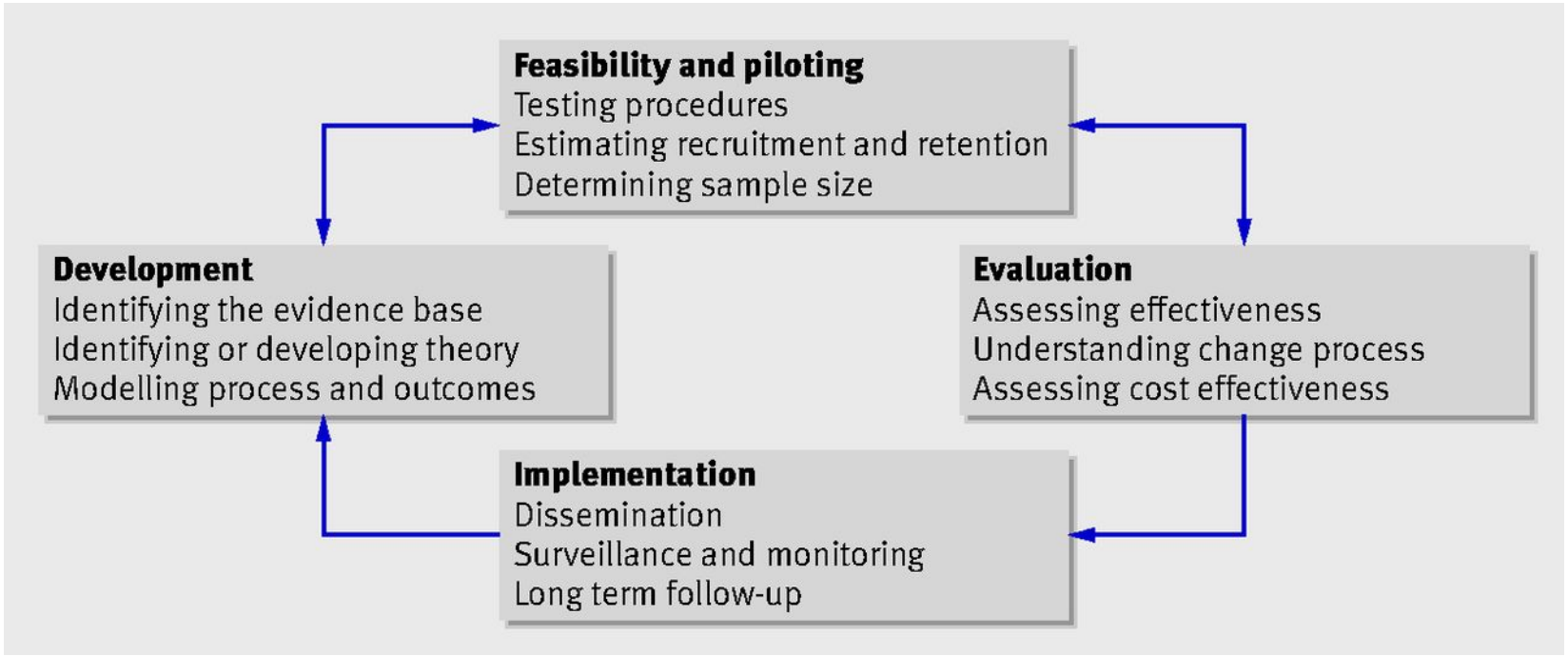
How might we do this?

- There are many interacting components
- There are a number of novel behaviours required by those delivering or receiving the intervention, and some of these changes are challenging
- Many groups and organisational levels are targeted by the intervention
- There is a wide range of possible outcome measures
- Interventions will likely need tailoring to local environment

Complex Interventions

"Best practice is to develop interventions systematically, using the best available practice and appropriate theory, then to test them using a carefully phased approach, starting with a series of pilot studies targeted at each of the key uncertainties in the design, and then moving on to an exploratory and then a definitive evaluation. The results should be disseminated as widely and as persuasively as possible, with further research to assist and monitor the process of implementation"

Peter Craig et al. BMJ 2008;337:bmj.a1655

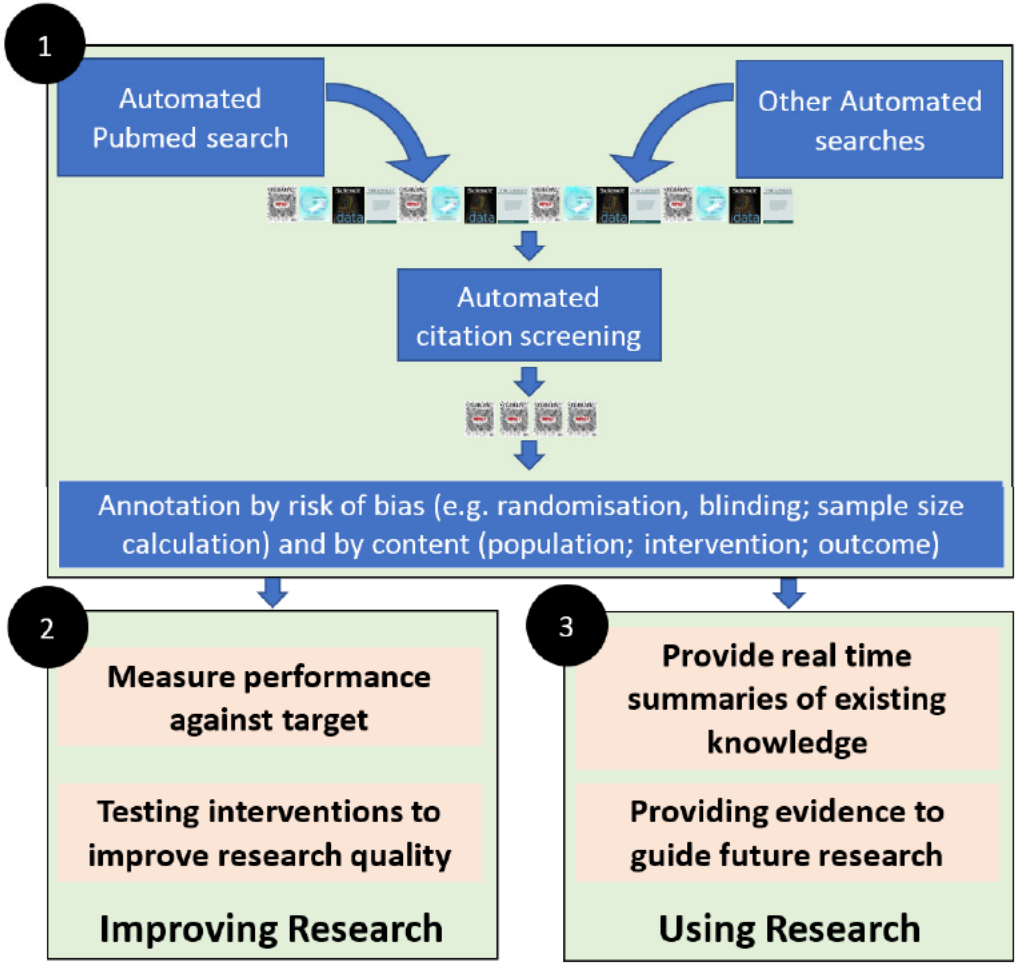


Peter Craig et al. BMJ 2008;337:bmj.a1655

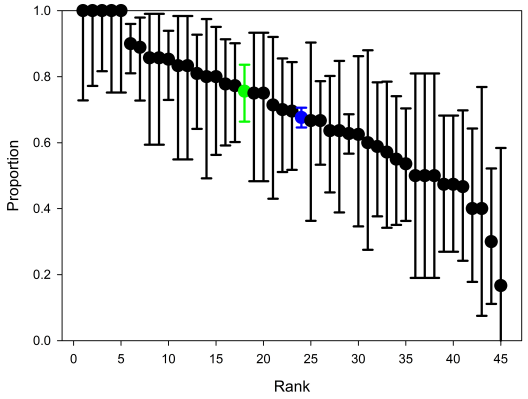
Possible approaches

Training	Undergraduate Post graduate Faculty	Mandatory or elective? As a requirement for conducting *in vivo* research?	Does certificated training to an enhanced level provide an advantage in grant application success rates?
Incentives	access to resources such as travel funds for junior staff, publication costs, near miss funding	• to pre-register study protocols • for BioRxiv publication • for Open Access publication • for making data and code availabe	
Support	Methodological support for review of experimental design prior to grant submission Review of manuscripts prior to journal submission Review of proposed research methodology at stage of requesting to animal house that procedure be permitted		
Admin	Can internal processes be made less burdensome while still achieving their purpose Can changes to institutional policies for appointment and promotion be used to encourage research improvement?		

Measuring outcomes



Blinding



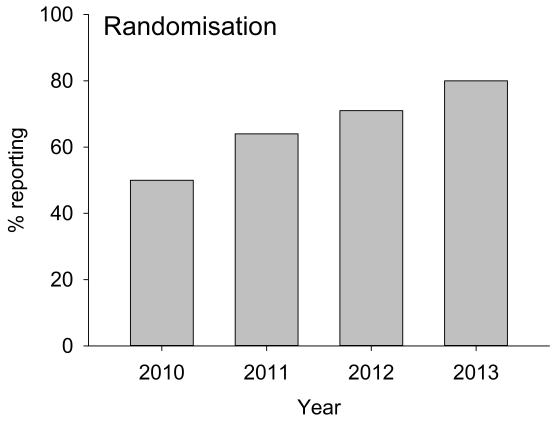
Research Improvement at Journals

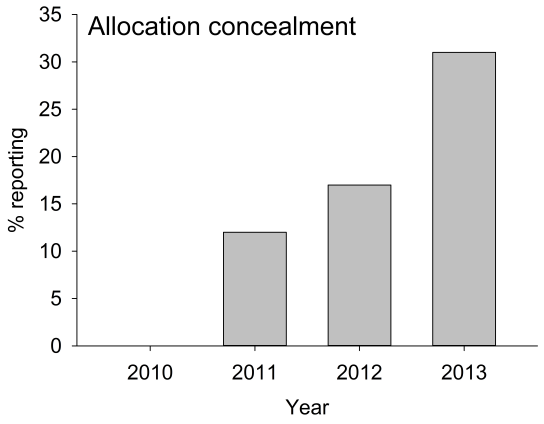
Comments, Opinions, and Reviews

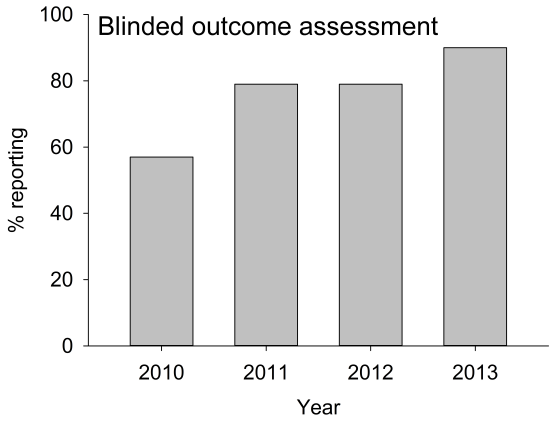
Good Laboratory Practice

Preventing Introduction of Bias at the Bench

Malcolm R. Macleod; Marc Fisher; Victoria O'Collins; Emily S. Sena; Ulrich Dirnagl; Philip M.W. Bath; Alistair Buchan; H. Bart van der Worp; Richard Traystman; Kazuo Minematsu; Geoffrey A. Donnan; David W. Howells







Minnerup et al, 2016

Ramirez et al Circ Res 2017

Supplemental Table: Comparison of study design element implementation in preclinical studies before and after the implementation the *Stroke* Basic Science Checklist, stratified by journal of publication

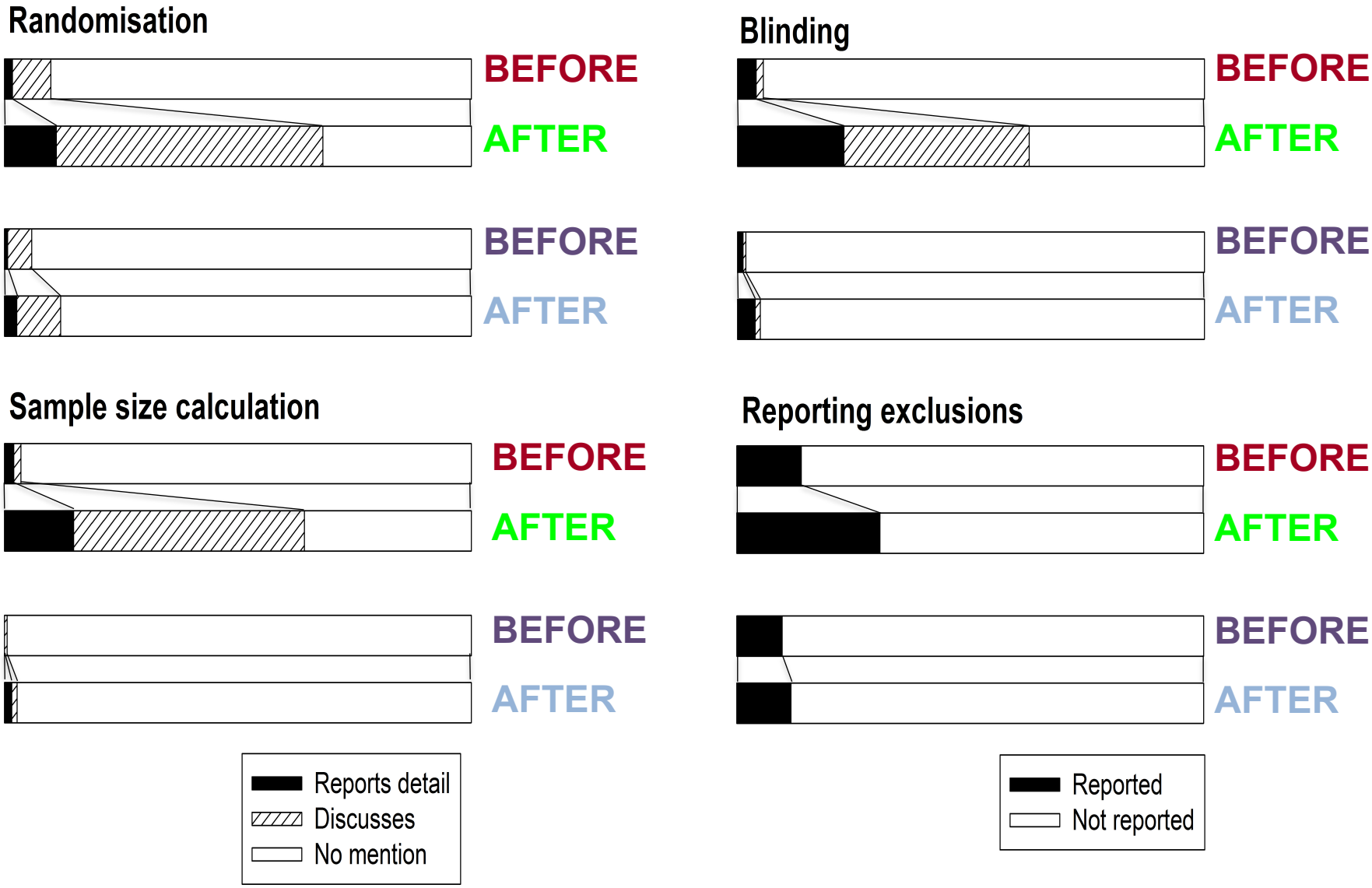
	Period 1* *n* (%)	Period 2* *n* (%)	Crude OR (95% CI)	*P*	Adjusted OR (95% CI)†	*P*†
Circulation	*n*=464	*n*=208				
Randomization	107 (23.1)	36 (17.3)	0.7 (0.5-1.1)	0.093	0.7 (0.4-1.1)	0.119
Blinding	169 (36.4)	59 (28.4)	0.7 (0.5-1.0)	0.042	0.7 (0.5-1.0)	0.043
Sample size estimation	7 (1.5)	5 (2.4)	1.6 (0.5-5.1)	0.422	NR	
Inclusion of both sexes	64 (13.8)	29 (13.9)	1.0 (0.6-1.6)	0.959	1.0 (0.6-1.6)	0.967
Circulation Research	*n*=303	*n*=183				
Randomization	35 (11.6)	29 (15.8)	1.4 (0.8-2.5)	0.176	1.4 (0.8-2.5)	0.261
Blinding	93 (30.7)	60 (32.8)	1.1 (0.7-1.6)	0.630	0.9 (0.6-1.4)	0.788
Sample size estimation	1 (0.3)	1 (0.3)	1.7 (0.1-26.7)	0.721	NR	
Inclusion of both sexes	57 (18.8)	33 (18.0)	0.9 (0.6-1.5)	0.830	1.0 (0.6-1.6)	0.937
Hypertension	*n*=485	*n*=375				
Randomization	104 (21.4)	81 (21.6)	1.0 (0.7-1.4)	0.956	1.2 (0.9-1.7)	0.298
Blinding	101 (20.8)	86 (22.9)	1.1 (0.8-1.6)	0.457	1.1 (0.8-1.5)	0.617
Sample size estimation	0 (0)	1 (0.3)	→∞ (0.0-∞)	0.946	NR	
Inclusion of both sexes	43 (8.9)	36 (9.6)	1.1 (0.7-1.7)	0.712	1.1 (0.7-1.7)	0.798
Stroke	*n*=316	*n*=185				
Randomization	120 (38.0)	119 (64.3)	2.9 (2.0-4.3)	<0.0001	3.2 (2.1-4.7)	<0.0001
Blinding	171 (54.1)	144 (77.8)	3.0 (2.0-4.5)	<0.0001	3.0 (2.0-4.5)	<0.0001
Sample size estimation	10 (3.2)	35 (18.9)	7.1 (3.4-14.8)	<0.0001	8.2 (3.7-18.4)	<0.0001
Inclusion of both sexes	15 (4.7)	20 (10.8)	2.4 (1.2-4.9)	0.012	2.4 (1.2-4.9)	<0.0001
ATVB	*n*=476	*n*=401				
Randomization	61 (12.8)	48 (12.0)	0.9 (0.6-1.4)	0.706	0.9 (0.6-1.4)	0.668
Blinding	130 (27.3)	97 (24.2)	0.8 (0.6-1.2)	0.293	0.7 (0.5-1.0)	0.026
Sample size estimation	2 (0.4)	10 (2.5)	6.1 (1.3-27.8)	0.021	NR	
Inclusion of both sexes	72 (15.1)	52 (13.0)	0.8 (0.6-1.2)	0.361	0.8 (0.6-1.3)	0.411

NR: not reported due to small number of events per predictor variable; OR: odds ratio

*Periods 1 and 2 correspond to before and after the date of implementation of the 'Basic Science Checklist' by *Stroke*, respectively

†Adjusted for cardiovascular disease studied and animal model used

Impact of NPG checklist



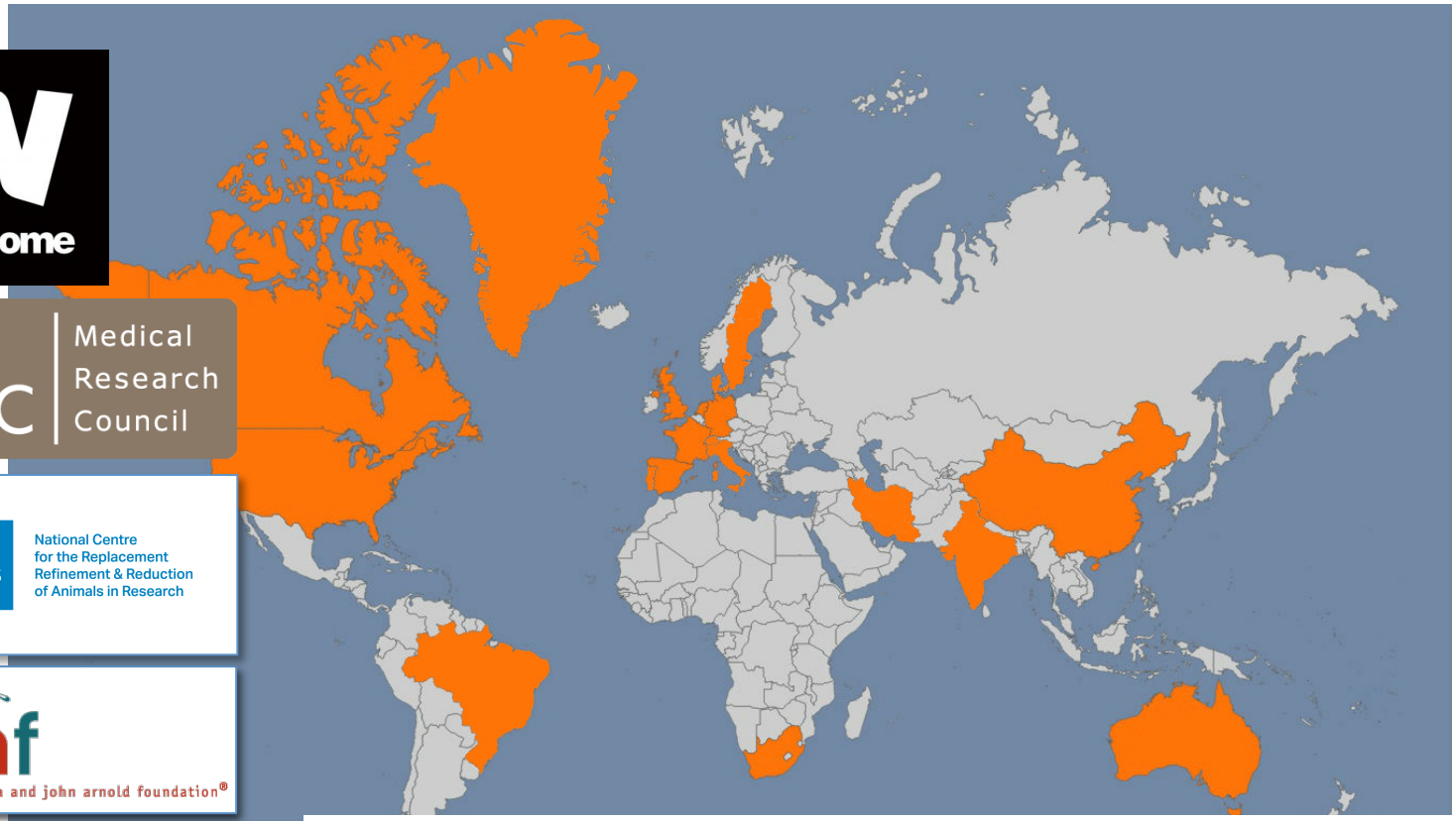
Recommendations

- Every institution which conducts research should have a formal Research Improvement Strategy
- Funders should consider the high returns on Replication studies (30-50% chance of revising existing knowledge)
- Pivotal research findings should be challenged in prospective multicenter replication studies prior to exploitation

Biomedical research investment

- $300bn globally, €50bn in Europe
- Glasziou and Chalmers claim 85% wasted
- Even if waste is only 50%, improvements which reduced that by 1% would free $3bn globally, €500m in Europe, every year.
- Investing ~1% of research expenditure in improvement activity would go a long way

If you are planning a systematic review or meta-analysis of animal data, CAMARADES are here to help: malcolm.macleod@ed.ac.uk



wellcome

MRC Medical Research Council

NC 3Rs National Centre for the Replacement Refinement & Reduction of Animals in Research

ljaf laura and john arnold foundation®

Newton Fund

efpia

imi innovative medicines initiative

The project leading to this application has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

CAMARADES: Bringing evidence to translational medicine