Study design, internal and external validity, explorative versus confirmatory studies

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
I am not in the office at the moment. Send any work to be translated.
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.
Treatment of experimental stroke with low-dose glutamate and homeopathic Arnica montana*

W. Jonas¹, Y. Lin², A. Williams², F. Tortella², R. Tuma³
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

- In vitro and in vivo: 1026
- Tested in vivo: 603
  - Effective in vivo: 374
  - Tested in clinical trial: 97
    - Effective in clinical trial: 1

O’ Collins et al, 2006

CAMARADES: Bringing evidence to translational medicine
“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
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.
Take 250 in vivo studies ...

**STATUS QUO:** Most studies have a statistical power of only 20% and a $P$ value of 0.05, meaning many more false findings (PPV of 50%). This reflects a sample size of about 10 mice per study.

- 10 promising molecules found
- 10 false positives found
- 40 undetected
- 190 true negative results (rarely published)

20 preclinical studies showed promise and were published, but 10 (50%) were false positives.
Camarades: Bringing evidence to translational medicine

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
• $p_1 \propto$
  – strength of original evidence
  – expert critical appraisal

• For each study, also know power of replication study, so can predict probability of successful replication ($=p_1 \times \text{power}$)

• Averaging across 41 studies,
  
  $p(\text{rep}) = 0.53$, $p(\text{non-rep}) = 0.47$

  $\therefore$ expected non-replication $= 19$ studies
  observed non-replication $= 25$ studies
  attributable non-replication $= 76\%$
Risk of bias in animal studies

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

Macleod et al, 2008
Regulation of REM and Non-REM Sleep by Periaqueductal GABAergic Neurons

Franz Weber¹,³, Johnny Phong Hoang Do¹, Shinjae Chung¹,³, Kevin T. Beier², Mike Bikov¹, Mohammad Saffari Doost¹ & Yang Dan¹

**Sample sizes.** For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions.
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”

<table>
<thead>
<tr>
<th>Group</th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Maze bright”</td>
<td>1.33</td>
<td>1.60</td>
<td>2.60</td>
<td>2.83</td>
<td>3.26</td>
</tr>
<tr>
<td>“Maze dull”</td>
<td>0.72</td>
<td>1.10</td>
<td>2.23</td>
<td>1.83</td>
<td>1.83</td>
</tr>
<tr>
<td>Δ</td>
<td>+0.60</td>
<td>+0.50</td>
<td>+0.37</td>
<td>+1.00</td>
<td>+1.43</td>
</tr>
</tbody>
</table>

Rosenthal and Fode (1963), Behav Sci 8, 183-9
Evidence from various neuroscience domains ...

Stroke

Alzheimer's disease

Multiple Sclerosis

Parkinson's disease

CAMARADES: Bringing evidence to translational medicine
The scale of the problem

RAE 1173

“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
Trans-lational research
Cis-lational research
If $\sum(\text{knowledge}) < \text{threshold}$ \Rightarrow cis - lation

If $\sum(\text{knowledge}) > \text{threshold}$ \Rightarrow trans - lation
Cumulative random effects meta-analysis of tPA in stroke

Precision of estimation of effect of species, a known variable of interest
Impact of known and latent variables:
As new studies are published (x-axis) then heterogeneity, estimated by Q or I^2 increases, and the variance of the overall effect s^2 falls. However, \( \tau^2 \) (the “between study variance”) shows a different pattern, with an initial peak, then a second rise, then falling again. We see this also in the IL-1 RA dataset. It may be that the second fall in \( \tau^2 \) corresponds to a dataset where possible heterogeneity has been adequately sampled.
A little bit of statistics

- **p-threshold** – the probability of observing an effect that big, or more extreme, if the null hypothesis is correct
  - Traditionally $p<0.05$
- **Power** – the probability of observing an effect of a given magnitude if it is present
## It’s a 2 x 2 table

<table>
<thead>
<tr>
<th></th>
<th>Test +ve</th>
<th>Test -ve</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truly +ve</strong></td>
<td>a</td>
<td>b</td>
<td><strong>Power = a/(a+b)</strong></td>
</tr>
<tr>
<td><strong>Actually -ve</strong></td>
<td>c</td>
<td>d</td>
<td><strong>p = c/(c+d) = 0.05</strong></td>
</tr>
<tr>
<td><strong>PPV = a/(a+c)</strong></td>
<td></td>
<td></td>
<td><strong>NPV = d/(b+d)</strong></td>
</tr>
<tr>
<td><strong>Prevalence = (a+b)/(c+d)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Take 250 in vivo studies ...
...with $p < 0.01$, power @ 80%

**Proposed Standards:** To achieve a PPV of 95%, study results would need a $P$ value of 0.01 and a large enough sample size to reach 80% statistical power (typically >75 mice per study).

- 40 promising molecules found
- 2 false positives found
- 10 undetected
- 198 true negative results

42 studies showed promise and were published, and only 2 (5%) were false positives.
The key to treating Alzheimer's could be to block the toxins produced by Porphyromonas gingivalis, the main bacteria in chronic gum disease. [link]

Advisory: this paper does NOT provide convincing evidence of this link (most of the work in mice, data from humans not compelling at present) and the study is carried out by US companies that sell blockers of AD.

Not just that; chances that the toxin from a bug causes AD - before this experiment - pretty low - say 1/10,000. Of 50,000 expts like this 49995 should be -ve. At p<0.05 5% (2500) will be false +ve. At power of 20% only 1 will be true +ve, so chance this is correct 1/2501 = 0.04%
• New prior = 0.0004
• Power = 90%
• Different p thresholds:
  – Posterior | sig@0.01 = 0.035 [alt 0.99996]
  – Posterior | sig@0.001 = 0.265 [alt 0.99996]
  – Posterior | sig@0.0001 = 0.783 [alt 0.99996]
Features of confirmatory experiments

• Clear *a priori* hypothesis
• Clear primary outcome measure
• Clear statistical analysis plan
• Well defined intervention
• Credible sample size calculation
  – What is the minimum effect size of interest?
  – Enough power to answer the question one way or another
Take things with a pinch of salt

• …unless you can be confident that primary outcome measure and statistical analysis plan were articulated before they saw the data
• … unless the outcome is clinically significant
• … if it’s a post hoc test
• … don’t rely on p<0.05
• … the more incredible the finding (ie the lower the prior), the less likely it is to be true
Researchers are different …

- F.F.P.
- HARKing
- Risks of bias
- Open Science
- Preregistration

number vs. quality

CAMARADES: Bringing evidence to translational medicine
Research Improvement Strategy
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