Statistical Problems with Preclinical Research- And How to Fix Them

Timothy T. Houle, PhD

Making Research Findings Less False

Timothy T. Houle, PhD

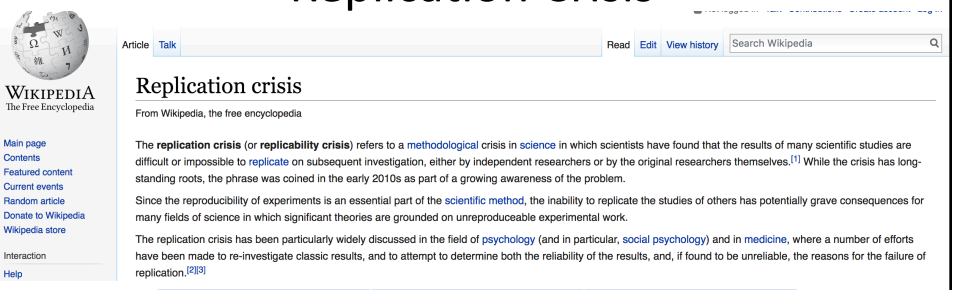
Disclosures

- NIH Grants
  - NS065257
  - GM113852
- Industry: None

Why Most Published Research Findings Are False
-Ioannidis (2005)

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Exhibit A: Replication Crisis



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help

Article Talk | Read Edit View history | Search Wikipedia

Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis**) refers to a methodological crisis in science in which scientists have found that the results of many scientific studies are difficult or impossible to replicate on subsequent investigation, either by independent researchers or by the original researchers themselves.[1] While the crisis has long-standing roots, the phrase was coined in the early 2010s as part of a growing awareness of the problem.

Since the reproducibility of experiments is an essential part of the scientific method, the inability to replicate the studies of others has potentially grave consequences for many fields of science in which significant theories are grounded on unreproduceable experimental work.

The replication crisis has been particularly widely discussed in the field of psychology (and in particular, social psychology) and in medicine, where a number of efforts have been made to re-investigate classic results, and to attempt to determine both the reliability of the results, and, if found to be unreliable, the reasons for the failure of replication.[2][3]

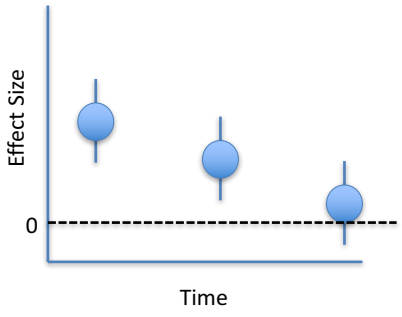
Fail to Replicate:	Others Work	Their Own Work
Chemistry	90%	60%
Biology	80%	60%
Physics	70%	50%
Medicine	70%	60%
Earth Science	60%	40%

Baker (2016). 1,500 scientists lift the lid on reproducibility

Exhibit B: The Mystery of the Disappearing Effect Size

Decline Effect (Rhine, 1930)

Generalizations Decay (1975)



Schooler, J. W.; Engstler-Schooler, T. Y. (1990). "Verbal overshadowing of visual memories: Some things are better left unsaid". Cognitive Psychology. 22 (1): 36–71.

Jonah_Lehrer (2010). "The Truth Wears Off". The New Yorker.

Exhibit C:
The Dog That Didn't Bark

- Prediction models that are not replicated, used, or applied in any way
- 1978 to 2016: 56,202 prediction models
  - 346 replication studies

Causes of the Crisis

- Multifaceted
  - Unprecedented rate of publication
  - Pressures to publish
  - Worship of novelty
  - Fraud
  - P-hacking
  - Researcher degrees of freedom
  - Selective Publication

Garden of Forking Paths
Gelman (2013)

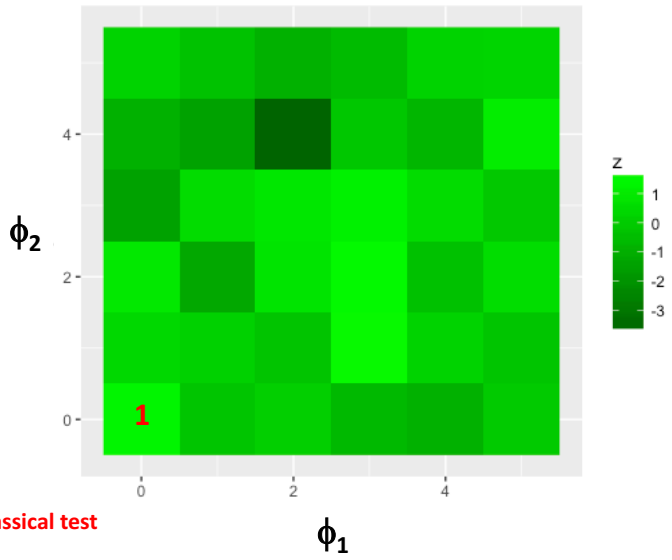
… The data must be a ***random***, representative sample from the process being studied

… The observed data represent one realization of a process that could be indefinitely repeated

Scenario	Test Statistic	
1. Simple classical test	T(y)	One planned statistical inference
2. Test pre-chosen from set of possible tests	T (y; ϕ)	One test with pre-registered ϕ
3. Test based on the data	T (y ; ϕ(y))	Only one test. Different test would have been performed given different data
4. Fishing	T (y ; ϕ_j)	Performing j tests and reporting the best one(s)

ϕ: control variables, covariates, transformations, data coding rules, exclusion, outliers, main effects, interactions, subgroups, alternate outcomes, direction of effect

Garden of Forking Paths

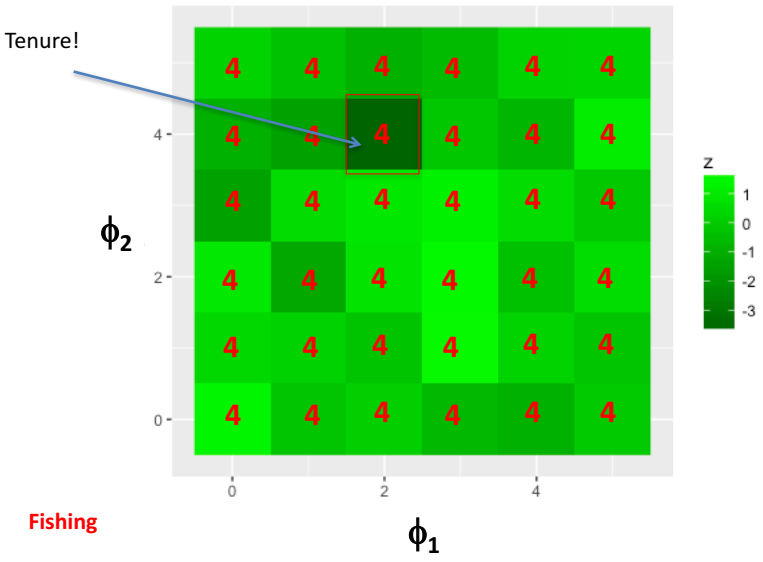


Simple classical test

Scenario	Test Statistic	
1. Simple classical test	T(y)	One planned statistical inference
2. Test pre-chosen from set of possible tests	T (y; ϕ)	One test with pre-registered ϕ
3. Test based on the data	T (y ; ϕ(y))	Only one test. Different test would have been performed given different data
4. Fishing	T (y ; ϕ_j)	Performing j tests and reporting the best one(s)

ϕ: control variables, covariates, transformations, data coding rules, exclusion, outliers, main effects, interactions, subgroups, alternate outcomes, direction of effect

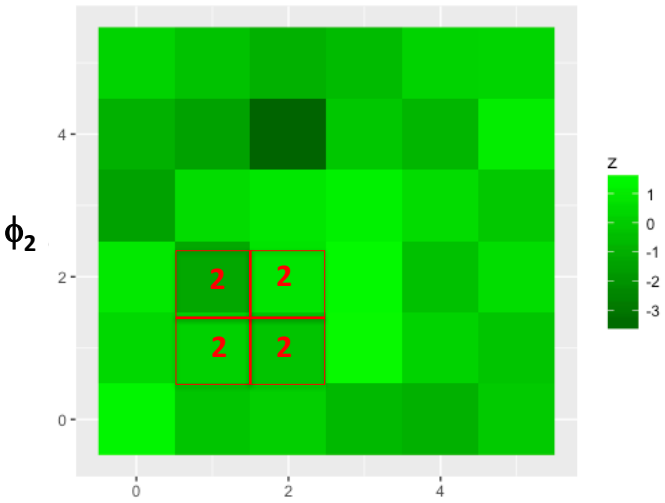
Garden of Forking Paths



Scenario	Test Statistic	
1. Simple classical test	T(y)	One planned statistical inference
2. Test pre-chosen from set of possible tests	T (y; ϕ)	One test with pre-registered ϕ
3. Test based on the data	T (y ; ϕ(y))	Only one test. Different test would have been performed given different data
4. Fishing	T (y ; ϕ_j)	Performing j tests and reporting the best one(s)

ϕ: control variables, covariates, transformations, data coding rules, exclusion, outliers, main effects, interactions, subgroups, alternate outcomes, direction of effect

Garden of Forking Paths

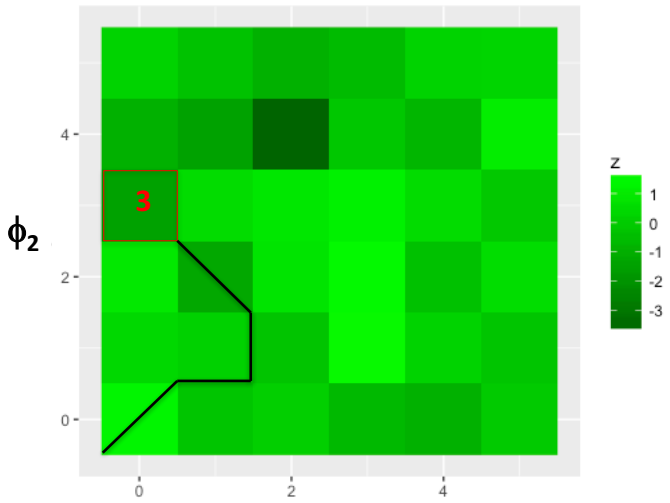


Test pre-chosen (registered) from many tests

Scenario	Test Statistic	
1. Simple classical test	T(y)	One planned statistical inference
2. Test pre-chosen from set of possible tests	T (y; ϕ)	One test with pre-registered ϕ
3. Test based on the data	T (y ; ϕ(y))	Only one test. Different test would have been performed given different data
4. Fishing	T (y ; ϕ_j)	Performing j tests and reporting the best one(s)

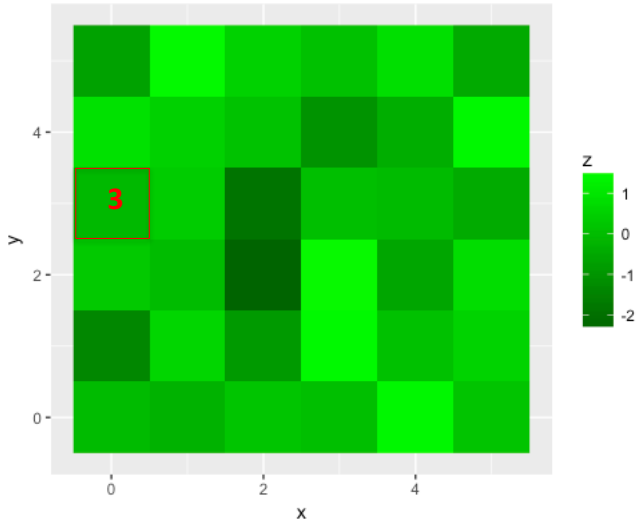
ϕ: control variables, covariates, transformations, data coding rules, exclusion, outliers, main effects, interactions, subgroups, alternate outcomes, direction of effect

Garden of Forking Paths

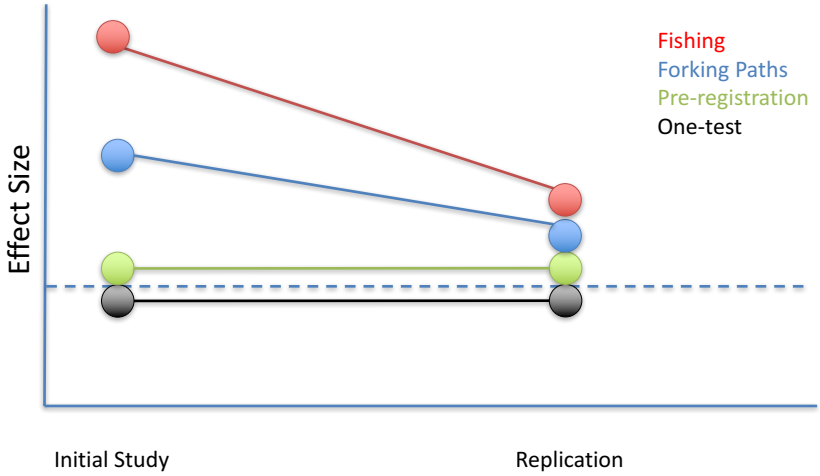


Test Based on data

Garden of Forking Paths



Replication Behavior



Scenario	Test Statistic	
1. Simple classical test	T(y)	One planned statistical inference
2. Test pre-chosen from set of possible tests	T (y; ϕ)	One test with pre-registered ϕ
✔ 3. Test based on the data	T (y ; ϕ(y))	Only one test. Different test would have been performed given different data
4. Fishing	T (y ; ϕ_j)	Performing j tests and reporting the best one(s)

ϕ: control variables, covariates, transformations, data coding rules, exclusion, outliers, main effects, interactions, subgroups, alternate outcomes, direction of effect

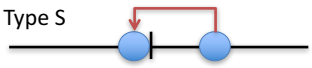
Definition of False

Traditional

- *Type I error*: Reject null hypothesis when it is true
- *Type II error*: Fail to reject null hypothesis when it is false

Gelman

- Type M error: Errors in the magnitude of the estimated effect size
- Type S error: Errors in the sign of the estimated effect size



http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.432.8657&rep=rep1&type=pdf

Recipe for Greater Chances of Being False

- Small effect size
- Small (modest) sample size
- Large measurement error
- High variation

- Low prior probability of effect

Top 3 Forking Paths

- One reasonable hypothesis maps on to several reasonable statistical hypotheses (i.e., one to many)
  - PONV is associated with age
- Statistical interaction (moderation) must be taken into consideration for the primary interpretation
  - Age x sex is needed to consider the effect of age
- Adding confounder control post hoc
  - We should control for several covariates as they seem to be confounding the age association

Possible Reactions

Deductive	Inductive
• This is a major concern	• Not concerned
• Science should proceed from carefully crafted inferences	• The very idea of science is to learn from data; you have to explore your data to know what it tells you.
• Few inferences, high confidence in them	• Many inferences are okay
• P-values (and CI) don't really indicate what they are supposed to under most applied circumstances	• P-values (and CI) are merely tools, I like to display the actual data anyway
• We should change what we do	• Carry on!

Recommendations:

- Avoid forking paths
  - Pre-registration
  - Fixed statistical analysis plans
    - Primary designation
    - Moderators
    - Multiplicity adjustments
  - Reproducible documents

https://ropensci.org/

Recommendations

- Okay, you insist on conducting data-driven analyses:
  - Be aware
    - P-values are suspect
    - CI coverage is too narrow
    - Effect sizes will regress to 0 (how much?)
  - Report
    - Describe the nature of the plan of analysis
    - Attempt to describe the researcher degrees of freedom
    - Provide the issue in the Discussion

Recommendations

- Okay, you insist on conducting data-driven analyses:
  - Formal inductive inference
    - Bayesian inference is inductive inference for adults
  - Allow others to reproduce your work
  - Internal validation
    - Bootstrapping, etc.

OPEN ACCESS Freely available online

PLoS MEDICINE

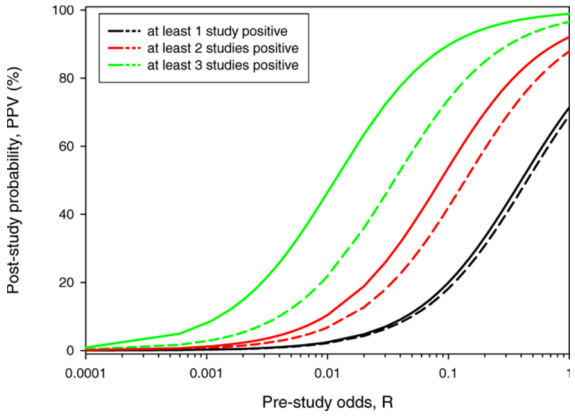
Essay

Most Published Research Findings Are False—But a Little Replication Goes a Long Way

Ramal Moonesinghe*, Muin J. Khoury, A. Cecile J. W. Janssens

We know there is a lot of lack of replication in research findings, most notably in the field of genetic associations [1–3]. For example, a survey of 600 positive associations between gene variants and common diseases showed that out of 166 reported associations studied three or more times, only six were replicated consistently [4]. Lack of replication results from a number of factors such as publication bias, selection bias, Type I errors, population stratification (the mixture of individuals from heterogeneous genetic backgrounds), and lack of statistical power [5].

In a recent article in *PLoS Medicine*, John Ioannidis quantified the theoretical basis for lack of replication by deriving the positive predictive value (PPV) of the truth of a research finding on the basis of a combination



Thank you!

- thoule1@mgh.harvard.edu