Plan

- Issues of reproducibility in science, historical perspective

Plan

- Issues of reproducibility in science, historical perspective - Was Ioannidis right?

Plan

- Issues of reproducibility in science, historical perspective - Was Ioannidis right? - Annecdotal evidence

Plan

- Issues of reproducibility in science, historical perspective - Was Ioannidis right? - Annecdotal evidence
- Where is the problem coming from?
  - computations, stats, sociology
  - cf everything matters

Plan

- Issues of reproducibility in science, historical perspective - Was Ioannidis right? - Annecdotal evidence
- Where is the problem coming from?
  - computations, stats, sociology
  - cf everything matters
- Emphasis on statistical issues

Plan

- Issues of reproducibility in science, historical perspective - Was Ioannidis right? - Annecdotal evidence
- Where is the problem coming from?
  - computations, stats, sociology
  - cf everything matters
- Emphasis on statistical issues
- Are there solutions ?

Issues of reproducibility in science

Issues of reproducibility in science

- Data bias and specificities

Issues of reproducibility in science

- Data bias and specificities
- Computational issues

Issues of reproducibility in science

- Data bias and specificities
- Computational issues
- Statistical: $p=0.05$ is weak evidence

Issues of reproducibility in science

- Data bias and specificities
- Computational issues
- Statistical: $p=0.05$ is weak evidence
- Ioannidis theoretical arguments "Why most research findings…"

Issues of reproducibility in science

- Data bias and specificities
- Computational issues
- Statistical: p=0.05 is weak evidence
- Ioannidis theoretical arguments "Why most research findings..."
- Statistics is about the **practice** of statistics

Issues of reproducibility in science

- Data bias and specificities
- Computational issues
- Statistical: $p=0.05$ is weak evidence
- Ioannidis theoretical arguments "Why most research findings..."
- Statistics is about the **practice** of statistics
- Publication **pressure** is still immense

Anecdotal evidence 1

Published in final edited form as:
JAMA Psychiatry. 2017 January 01; 74(1): 47–55. doi:10.1001/jamapsychiatry.2016.2783.

Altered Brain Activity in Unipolar Depression Revisited Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD, **Edna C. Cieslik, PhD**, **Ilinca Serbanescu, MSc**, **Angela R. Laird, PhD**, **Peter T. Fox, MD**, and **Simon B. Eickhoff, MD**

During the past 20 years, numerous neuroimaging experiments have investigated aberrant brain activation during cognitive and emotional processing in patients with unipolar depression.

Anecdotal evidence 1

HHS Public Access

Author manuscript

JAMA Psychiatry. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

JAMA Psychiatry. 2017 January 01; 74(1): 47–55. doi:10.1001/jamapsychiatry.2016.2783.

Altered Brain Activity in Unipolar Depression Revisited Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD, **Edna C. Cieslik, PhD**, **Ilinca Serbanescu, MSc**, **Angela R. Laird, PhD**, **Peter T. Fox, MD**, and **Simon B. Eickhoff, MD**

During the past 20 years, numerous neuroimaging experiments have investigated aberrant brain activation during cognitive and emotional processing in patients with unipolar depression.

Anecdotal evidence 1

HHS Public Access

Author manuscript

JAMA Psychiatry. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

JAMA Psychiatry. 2017 January 01; 74(1): 47–55. doi:10.1001/jamapsychiatry.2016.2783.

Altered Brain Activity in Unipolar Depression Revisited Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD, **Edna C. Cieslik, PhD**, **Ilinca Serbanescu, MSc**, **Angela R. Laird, PhD**, **Peter T. Fox, MD**, and **Simon B. Eickhoff, MD**

> *During the past 20 years, numerous neuroimaging experiments have investigated aberrant brain activation during cognitive and emotional processing in patients with unipolar depression.*

$>$ In total, 57 studies with 99 individual neuroimaging experiments comprising in total 1058 patients were included; 34 of them tested cognitive and 65 emotional processing. Overall analyses across cognitive processing experiments ($P > .29$) and across emotional processing experiments ($P > .47$) revealed **no significant results.**

Anecdotal evidence 2: All foods cause cancer ? Schoenfeld 2013

- Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased (n = 103) or a decreased (n = 88) risk;

Anecdotal evidence 2: All foods cause cancer ? Schoenfeld 2013

- Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased ($n = 103$) or a decreased ($n = 88$) risk;
- 75% of the risk estimates had weak ($0.05 > P > 0.001$) or no statistical ($P > 0.05$) significance.

Anecdotal evidence 2: All foods cause cancer ? Schoenfeld 2013

- Of 264 single-study assessments, 191 (72%) concluded that the tested food was associated with an increased ($n = 103$) or a decreased ($n = 88$) risk;
- 75% of the risk estimates had weak ($0.05 > P > 0.001$) or no statistical ($P > 0.05$) significance.
- Meta-analyses presented more conservative results; only 13 (26%) reported an increased ($n = 4$) or a decreased ($n = 9$) risk

Computational problems

- OS can be a problem (same container, different segementation)
  - Glatard et al, 2015

Computational problems

- OS can be a problem (same container, different segementation)
  - Glatard et al, 2015
- Version of librairies and software

Computational problems

- OS can be a problem (same container, different segementation)
  - Glatard et al, 2015
- Version of librairies and software
- Algorithms initialization

Computational problems

- OS can be a problem (same container, different segementation)
  - Glatard et al, 2015
- Version of librairies and software
- Algorithms initialization
- Algorithms sensitivity to noise (Kiar et al)

Computational problems

- OS can be a problem (same container, different segementation)
    - Glatard et al, 2015
- Version of librairies and software
- Algorithms initialization
- Algorithms sensitivity to noise (Kiar et al)
- Software variation

Evil p-values: Significance testing as perverse probabilistic reasoning

Consider a typical medical research study, for example designed to test the efficacy of a drug, in which a null hypothesis H_0 ('no effect') is tested against an alternative hypothesis H_1 ('some effect'). Suppose that the study results pass a test of statistical significance (that is P-value <0.05) in favor of H_1. What has been shown?

1. H_0 is false.
2. H_1 is true.
3. H_0 is probably false.
4. H_1 is probably true.
5. Both (1) and (2).
6. Both (3) and (4).
7. None of the above.

Significance testing as perverse probabilistic reasoning

Table 1 Quiz answer profile

Answer	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number	8	0	58	37	6	69	12
Percent	4.2	0	30.5	19.5	3.2	36.3	6.3

- Westover, 2014

P-value Definition

Probability of observing a statistic equal to the one seen in the data, or one that is more “extreme”, when the null hypothesis is true

P-value requires:

- Knowledge of the null hypothesis

P-value requires:

- Knowledge of the null hypothesis
- Choice of a statistic

P-value requires:

- Knowledge of the null hypothesis
- Choice of a statistic
- Concept of repeating the whole study in the same way
    - Same study design

P-value requires:

- Knowledge of the null hypothesis
- Choice of a statistic
- Concept of repeating the whole study in the same way
    - Same study design
    - Same sampling scheme

P-value requires:

- Knowledge of the null hypothesis
- Choice of a statistic
- Concept of repeating the whole study in the same way
  - Same study design
  - Same sampling scheme
  - Same definition of the statistic

What happens if ... p is "significant" but study power is low ?

- Power : the probability of finding a significant p-value under H1

What happens if ... p is "significant" but study power is low ?

- Power : the probability of finding a significant p-value under H1
- Study in Button et al, 2013, more than half of the studies have less than 30% power

What happens if ... p is "significant" but study power is low ?

- Power : the probability of finding a significant p-value under H1
- Study in Button et al, 2013, more than half of the studies have less than 30% power
- Low Positive Predictive Value P(H_A true | test significant)

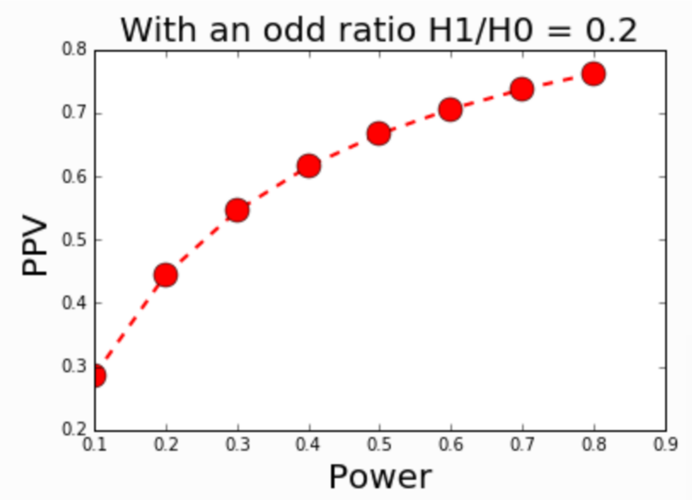
What happens if ... p is "significant" but study power is low ?

- Power : the probability of finding a significant p-value under H1
- Study in Button et al, 2013, more than half of the studies have less than 30% power
- Low Positive Predictive Value P(H_A true | test significant)
- Inflated effect size

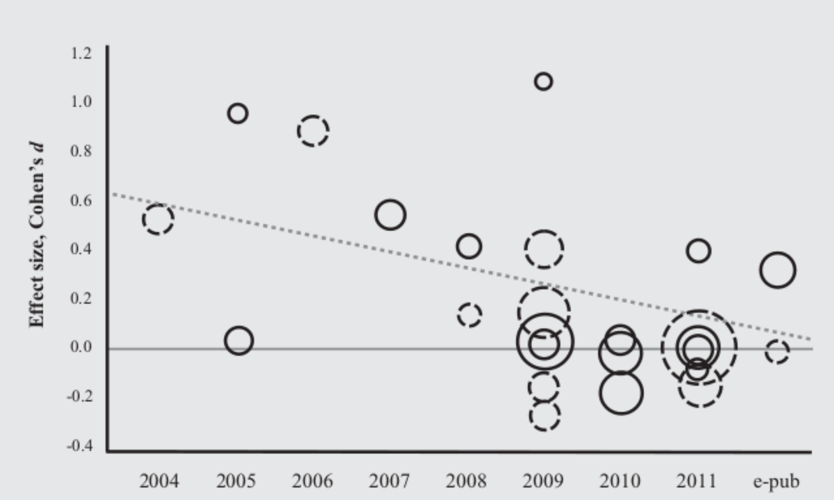
What happens if ... p is "significant" but study power is low ?

- Power : the probability of finding a significant p-value under H1
- Study in Button et al, 2013, more than half of the studies have less than 30% power
- Low Positive Predictive Value P(H_A true | test significant)
- Inflated effect size
- Depends on the prior probability of H_A and H_0

Low Positive Predictive Value : P(H_A is true | test is significant)



Inflated effect size Effect-size = f(years, sample, . . .)

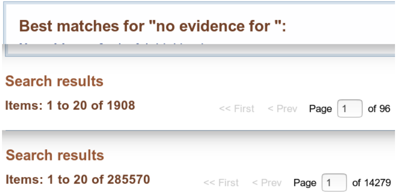


Molendijk, 2012, BDNF and hippocampal volume

What happens if . . . p is not significant? File drawer effect

- Described by Rosenthal in **1979**
- Most publications accepted if p<.05
- Hard to publish null results

“. . . whether you would be able to review the manuscript”No Evidence for an Effect of XXX on Hippocampal Volume in a YYY Sample“, by some-authors, submited for consideration in . . . ”



Are we always testing/publishing at p=0.05 ? Incentive perversion

- Implies P-Hacking and Harking
  - Simmons and Simonsohn 2011, P-curves

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables (r = .50)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Is p-hacking really happening ?

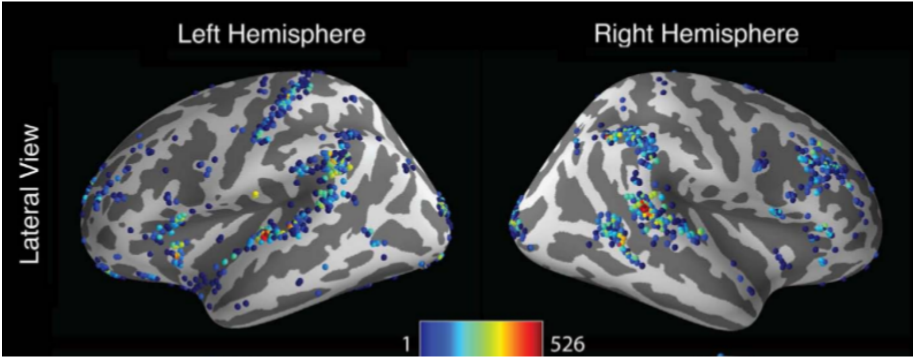
Open Access Research

BMJ Open **Identifying bioethical issues in biostatistical consulting: findings from a US national pilot survey of biostatisticians**

Min Qi Wang,[1] Alice F Yan,[2] Ralph V Katz[3]

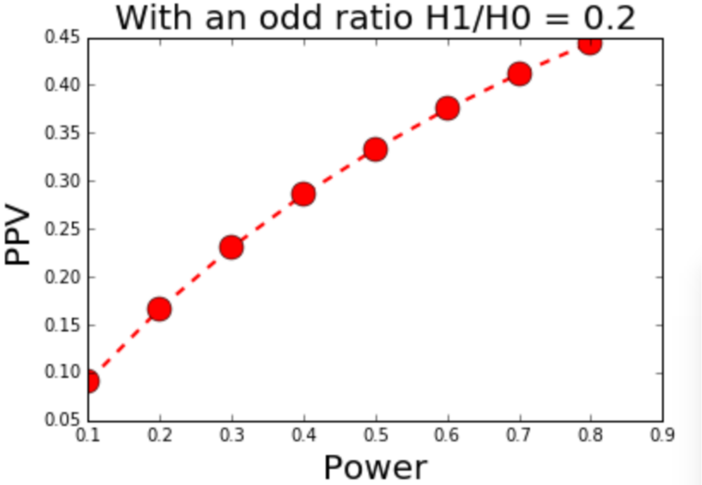
- study gives **clear evidence** that researchers make requests of their biostatistical consultants that are not only rated as **severe violations**, but further that these requests occur quite **frequently**.
- P-curve: Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014.
  - Principle: literature should not have that many p close to .05
  - p-values are uniformely distributed (how do you show that?)

Are we always testing/publishing at p=0.05 ? Incentive perversion

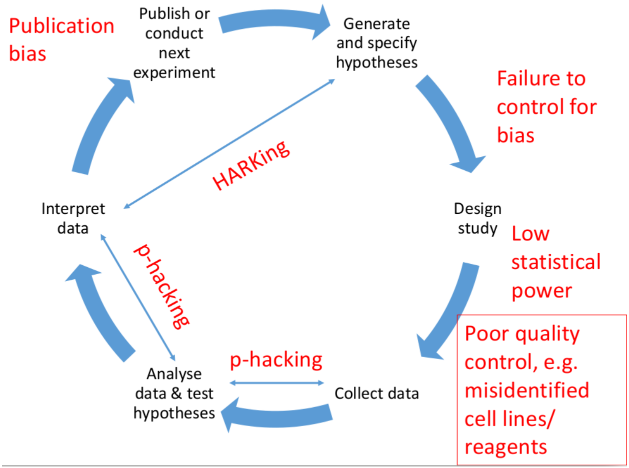


- Carp 2012

Low Positive Predictive Value : P(H_A is true | test is significant)



A possibly quite dire situation



- D. Bishop 2015

Solutions : sociological

- Ban p-values sounds a little extreme (BASP)
    - Btw: Nature editorial stated :
      "The closer to zero the P value gets, the greater the chance the null hypothesis is false."

Solutions : sociological

- Ban p-values sounds a little extreme (BASP)
  - Btw: Nature editorial stated :
    "The closer to zero the P value gets, the greater the chance the null hypothesis is false."
- Registered Reports
  - Seems a good solution in many cases: can implement a culture shift: worth the paper work !

Solutions : sociological

- Ban p-values sounds a little extreme (BASP)
  - Btw: Nature editorial stated :
    "The closer to zero the P value gets, the greater the chance the null hypothesis is false."
- Registered Reports
  - Seems a good solution in many cases: can implement a culture shift: worth the paper work !
- Cobidas and reporting best practices
  - community education and publishing efforts
  - standards for easing reuse of data (INCF, BIDS)
  - Long list of checkboxes in nature publications - Cobidas
  - Nature statistician review

Solutions : Technical

- Redefine significance

Solutions : Technical

- Redefine significance
- Use Bayesian framework ?

Solutions : Technical

- Redefine significance
- Use Bayesian framework ?
- Use Prediction framework ?

Solutions : Technical

- Redefine significance
- Use Bayesian framework ?
- Use Prediction framework ?

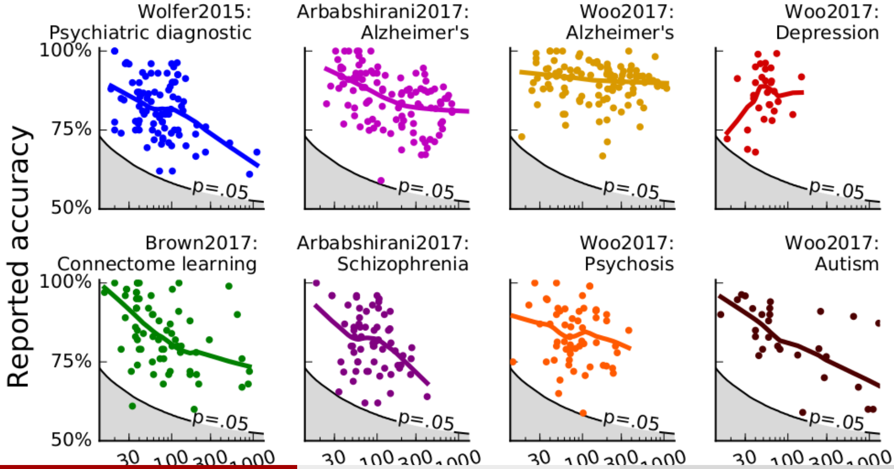
Conclusion: Is machine learning (prediction / classification) going to save us?

- Yes: Why ?

Conclusion: Is machine learning (prediction / classification) going to save us?

- Yes: Why ?
- No: Why ?

Is machine learning (prediction / classification) going to save us?



Conclusion: rephrase reproducibility into generalizability

- What do I generalize on ?
  - datasets ?
  - software ?
  - algorithms ? - initializations ?
  - populations ? ...
- Where is the biggest variation ?

Conclusion: Ioannidis again

- Young fields tend to have less stringent criteria
- Ioannidis 2005: When are results more likely to be false?
  - The smaller the studies ...
  - The smaller the effect size ...
  - The larger the number of tests ...
  - The more flexibility in the analyses
  - The more trendy ...
  - The more financial interest ...

Acknowledgements

- Repronim: D. Kennedy, S. Ghosh, Y. Halchenko, D. Keator, D. Jarecka, J. Grethe, M. Martone, etc. . .
- McGill: Peer Herholz, Lex Hutton, Celia Greenwood, Bettina Kemme, Samir Das, Shawn Brown, Alan Evans, Bratislav Misic
- Berkeley: M. D'Esposito, M. Brett, S. Van der Walt, J.Millman
- Pasteur: G. Dumas, R. Toro, T. Bourgeron, A. Beggiato
- Neurospin: B. Thirion, G. Varoquaux, V. Frouin, others

- **Hiring on reproducibility and neuroinformatics projects !**

Thank you for your attention - Questions ?