

The problems associated with the use of p-values in brain imaging and their effects on reproducibility

A reproducibility perspective

JB Poline

McGill University, UC Berkeley

June 17 2018

Plan

- Definition
- A quick historical perspective

Plan

- Definition
- A quick historical perspective
- Was Ioannidis right?

Plan

- Definition
- A quick historical perspective
- Was Ioannidis right?
- What is/are the problems?
 - technical
 - sociological

Plan

- Definition
- A quick historical perspective
- Was Ioannidis right?
- What is/are the problems?
 - technical
 - sociological
- Is there a solution?

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

Requires:

- Knowledge of the null hypothesis

Requires:

- Knowledge of the null hypothesis
- Choice of a statistic

Requires:

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

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

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

Issues of reproducibility in science

Credibility Crisis

The collage features several key articles and announcements:

- Los Angeles Times**: "Science has lost its way, at a big cost to humanity". Sub-headline: "Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true."
- Science** (AAAS): "Reproducibility". By Marcia McNitt. Sub-headline: "Science advances on a foundation of trusted data, but a new approach that scientists use to gain confidence in their results is being shaken by reports that a trust not reproducible. Because confidence in result community, we are announcing new initiatives Science. For preclinical studies (one of the target recommendations of the U.S. National Institute increasing transparency." Authors will indicate handling (such as how to deal with outliers), we ensure a sufficient signal-to-noise ratio, whether experimenter was blind to the conduct of the guidelines."
- The Scientist**: "NIH Tackles Irreproducibility". Sub-headline: "The federal agency speaks out about how to improve the quality of scientific research." By Jef Akst | January 28, 2014.
- Nature**: "Announcement: Reducing our irreproducibility". 24 April 2013. Sub-headline: "Over the past year, Nature has published a string of articles that reliability and reproducibility of published research (collected as..."
- Nature**: "Must try harder". Nature 483, 509 (29 March 2012) | doi:10.1038/483509a. Published online 28 March 2012. Sub-headline: "Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves."
- The Economist**: "Washington's larger surplus". Sub-headline: "How to do a nuclear deal with Iran. Investment tips from Nobel economists. Junk bonds are back. The meaning of Sachin Tendulkar."
- HOW SCIENCE GOES WRONG**: A graphic with the text "HOW SCIENCE GOES WRONG" and a small illustration of a globe.

How much is reproducibility impacted by our current statistical procedures ?

- Reproducibility a *computational* issue, Replicability a *statistical* issue

How much is reproducibility impacted by our current statistical procedures ?

- Reproducibility a *computational* issue, Replicability a *statistical issue*
- Evidence that this might be the case:
 - Works from statisticians to show that $p=0.05$ is weak evidence
 - Ioannidis theoretical arguments “Why most research findings. . .”

How much is reproducibility impacted by our current statistical procedures ?

- Reproducibility a *computational* issue, Replicability a *statistical issue*
- Evidence that this might be the case:
 - Works from statisticians to show that $p=0.05$ is weak evidence
 - Ioannidis theoretical arguments “Why most research findings. . .”
- Statistics is about the **practice** of statistics

How much is reproducibility impacted by our current statistical procedures ?

- Reproducibility a *computational* issue, Replicability a *statistical* issue
- Evidence that this might be the case:
 - Works from statisticians to show that $p=0.05$ is weak evidence
 - Ioannidis theoretical arguments “Why most research findings. . .”
- Statistics is about the **practice** of statistics
- Study on the statistical practices
 - Simmons and Simonsohn in psychology
 - Wang et al., 2018 in biomedical research

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

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, Ilina 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, Ilina 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

Historical perspective

- When did we start to talk of the problem?
 - almost as soon as the time of the p-value was defined, 1925

Historical perspective

- When did we start to talk of the problem?
 - almost as soon as the time of the p-value was defined, 1925
- Fisher conception:
 - an indication of something about data under H_0

Historical perspective

- When did we start to talk of the problem?
 - almost as soon as the time of the p-value was defined, 1925
- Fisher conception:
 - an indication of something about data under H_0
- Neyman-Pearson conception:
 - a decision making rule

Historical perspective

- When did we start to talk of the problem?
 - almost as soon as the time of the p-value was defined, 1925
- Fisher conception:
 - an indication of something about data under H_0
- Neyman-Pearson conception:
 - a decision making rule
- Which one is used today ?

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

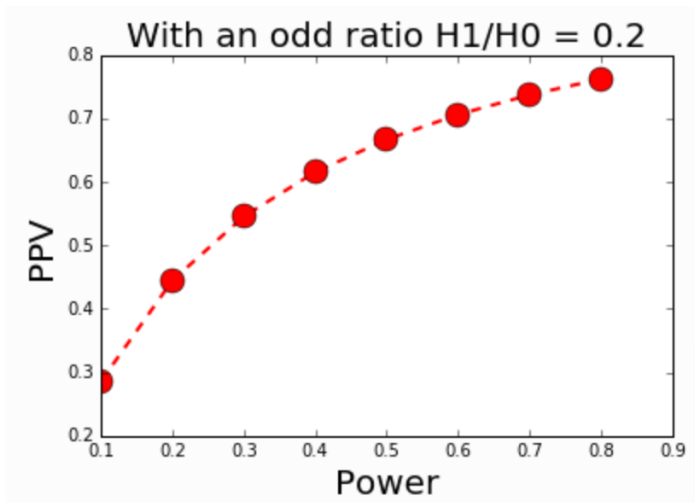
Answer	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number	8	0	58	37	6	69	12
Percent	4.2	0	30.5	19.5	3.2	36.3	6.3

- Westover, 2014

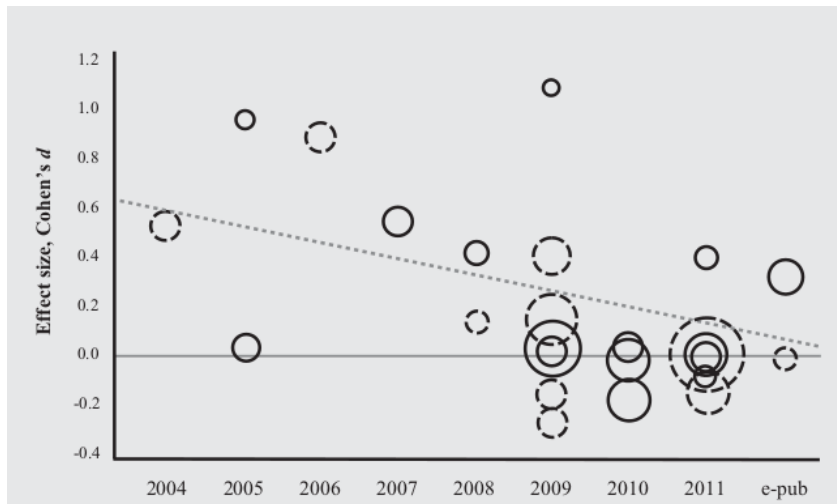
What happens if ... p is “significant” but study power is low ?

- Study in Button et al, 2013, more than half of the studies have less than 30% power
- Low Positive Predictive Value $P(H_A \text{ true} \mid \text{test significant})$
- Inflated effect size
- Depends on the prior probability of H_A and H_0

Low Positive Predictive Value : $P(H_A \text{ is true} \mid \text{test is significant})$



Inflated effect size Effect-size = $f(\text{years, sample, } \dots)$



Not everybody believes in power

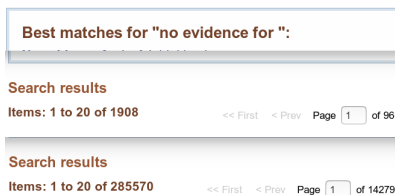
- Grant reviewer quote (grant on power rejected)

“... I am skeptical that searches of existing studies have information that's relevant and targeted enough to assessing power or reproducibility for scientifically interesting new designs.”

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

- Described first by Rosenthal in **1979**
- Most publications accepted only with $p < .05$
- 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, submitted for consideration in ..."



Wait - are we always testing/publishing at $p=0.05$?

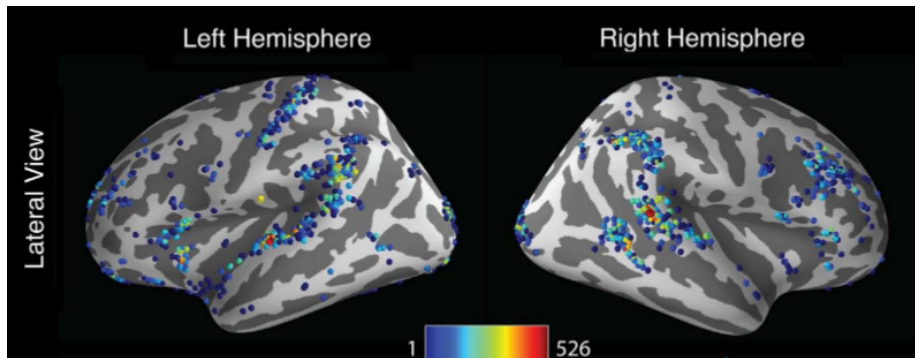
Incentive perversion

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

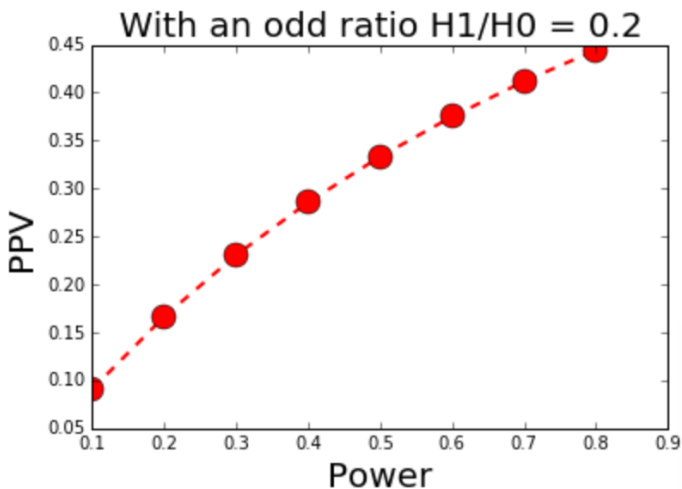
Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
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%

Wait - are we always testing/publishing at $p=0.05$?
Incentive perversion



Low Positive Predictive Value : $P(H_A \text{ is true} \mid \text{test is significant})$



Is this really happening ?

Open Access

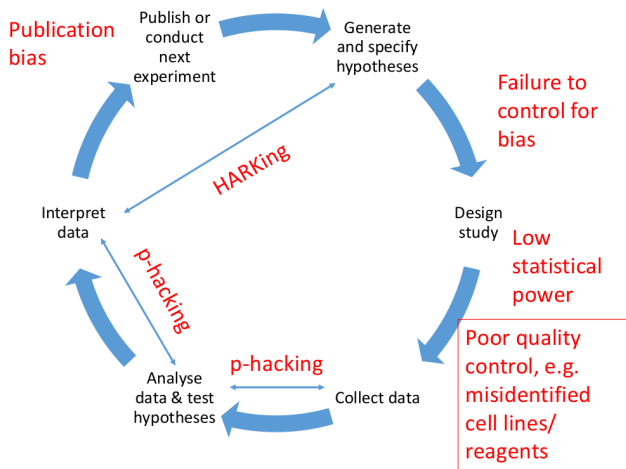
Research

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 Katz³

- 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**.

A possibly quite dire situation



- D. Bishop 2015

The reactions - the solutions?

- Technical:
 - Redefine significance
 - Use Bayesian framework
 - Prediction framework

The reactions - the solutions?

- Technical:
 - Redefine significance
 - Use Bayesian framework
 - Prediction framework
- Social: work with the journals
 - Ban p-values
 - Long list of checkboxes in nature publications - Cobidas
 - Nature statistician review
 - Registered Reports

Solutions - technical - redefine significance

- 70 prominent scientists worked on a google document ...

"We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005."

- move BF from **weak 2-3** to **strong 12-26 evidence** (under many H1)

Daniel J. Benjamin 1 *, James O. Berger 2 , Magnus Johannesson 3 *, Brian A. Nosek 4, 5 , E. - J. Wagenmakers 6 , Richard Berk 7, 10 , Kenneth A. Bollen 8 , Björn Brembs 9 , Lawrence Brown 10 , Colin Camerer 11 , David Cesarini 12, 13 , Christopher D. Chambers 14 , Merlise Clyde 2 , Thomas D. Cook 15,16 , Paul De Boeck 17 , Zoltan Dienes 18 , Anna Dreber 3 , Kenny Easwaran 19 , Charles Efferson 20 , Ernst Fehr 21 , Fiona Fidler 22 , Andy P. Field 18 , Malcolm Forster 23 , Edward I. George 10 , Richard Gonzalez 24 , Steven Goodman 25 , Edwin Green 26 , Donald P. Green 27 , Anthony Greenwald 28 , Jarrod D. Hadfield 29 , Larry V. Hedges 30 , Leonhard Held 31 , Teck Hua Ho 32 , Herbert Hoijtink 33 , James Holland Jones 39,40 , Daniel J. Hruschka 34 , Kosuke Imai 35 , Guido Imbens 36 , John P.A. Ioannidis 37 , Minjeong Jeon 38 , Michael Kirchler 41 , David Laibson 42 , John List 43 , Roderick Little 44 , Arthur Lupia 45 , Edouard Machery 46 , Scott E. Maxwell 47 , Michael McCarthy 48 , Don Moore 49 , Stephen L. Morgan 50 , Marcus Munafó 51, 52 , Shinichi Nakagawa 53 , Brendan Nyhan 54 , Timothy H. Parker 55 , Luis Pericchi 56 , Marco Perugini 57 , Jeff Rouder 58 , Judith Rousseau 59 , Victoria Savalei 60 , Felix D. Schönbrodt 61 , Thomas Sellke 62 , Betsy Sinclair 63 , Dustin Tingley 64 , Trisha Van Zandt 65 , Simine Vazire 66 , Duncan J. Watts 67 , Christopher Winship 68 , Robert L. Wolpert 2 , Yu Xie 69 , Cristobal Young 70 , Jonathan Zinman 71 , Valen E. Johnson 72 *

Solutions: is it a solution, really?

- 88 non less prominent scientists declare that this is not a solution !

Abstract: In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently *report and justify* all choices they make when designing a study, *including the alpha level*.

Solutions: is it a solution, really?

- 88 non less prominent scientists declare that this is not a solution !

Abstract: In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently *report and justify* all choices they make when designing a study, *including the alpha level*.

- results depend on power and prior
- results depend on H_1
- priors are really hard to estimate
- may make science more costly and analyses lose sensitivity

Original authors fight back !



BAYESIAN SPECTACLES

Powered by JASP

[Home](#) [About »](#) [JASP](#) [Artwork »](#) [Contact](#)

Redefine Statistical Significance Part VIII: How 88 Authors Overlooked a Giraffe and Sailed Straight into an Iceberg

POSTED ON SEP 19TH, 2017



The key point of the paper "Redefine Statistical Significance" is that p-just-below-.05 results should be approached with care. They should perhaps evoke curiosity, but they should *not* receive the blanket endorsement that is implicit in the bold claim "we reject the null hypothesis". The statistical argument is straightforward and has been known for over half a century: for p-just-below-.05 results, the alternative hypothesis does not convincingly outpredict the null hypothesis, not even when we *cheat* and cherry-pick the alternative hypothesis that is inspired by the data.

The claim that p-just-below-.05 results are evidentially weak was recently echoed by the *American Statistical Association* when they stated that "a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis" (Wasserstein and Lazar, 2016, p.



Bayesian Spectacles is powered by JASP: a free, friendly, and flexible software package for conducting statistical analyses. Discover JASP at jasp-stats.org and follow JASP on Twitter.

◊ General

Registered reports



Chris Chambers ✓ @chrisdc77 · May 28

Yes this sums up the view of many skeptical profs I've talked to who don't believe p-hacking /HARKing is a problem, that irreproducibility concerns are overblown & that good researchers are immune to biased reasoning. In this world, RRs are a solution looking for a problem.

Registered reports



Chris Chambers ✓ @chrisdc77 · May 28

Yes this sums up the view of many skeptical profs I've talked to who don't believe p-hacking /HARKing is a problem, that irreproducibility concerns are overblown & that good researchers are immune to biased reasoning. In this



Jack Gallant @gallantlab · May 28

As I said, I hate paperwork. If people are using proper methods pre-registration is unnecessary. The focus should be on how change experimental methods to facilitate construction of better, more replicable, generalizable quantitative models.

Registered reports



Chris Chambers ✓ @chrisdc77 · May 28

Yes this sums up the view of many skeptical profs I've talked to who don't believe p-hacking /HARKing is a problem, that irreproducibility concerns are overblown & that good researchers are immune to biased reasoning. In this



Jack Gallant @gallantlab · May 28

As I said, I hate paperwork. If people are using proper methods pre-registration is unnecessary. The focus should be on how change experimental methods to facilitate construction of better, more replicable, generalizable quantitative models.



Jack Gallant @gallantlab · May 28



RR seems to be primarily about reducing Type I error. But if you view PNHT as insufficient, merely a weak, poorly reasoned pretest of data quality, then it becomes obvious that the focus should be elsewhere. We need a revolution, not more paperwork.

Registered reports

Registered reports



Jack Gallant @gallantlab · May 28



For example, require effect size reports and demand greater evidence for small effects. Require people to report what proportion of individuals show the effect. Separate fit and test sets. Do generalization tests. Test quantitative predictions.

Registered reports



Jack Gallant @gallantlab · May 28



For example, require effect size reports and demand greater evidence for small effects. Require people to report what proportion of individuals show the effect. Separate fit and test sets. Do generalization tests. Test quantitative predictions.



Tal Yarkoni @talyarkoni · May 28



if anything, it should be much **easier** to criticize studies for using NHST when an RR is first submitted for review, than to wait until after the authors are happily trumpeting their $p < .001$ result and can say "but look, it's strong!"

Registered reports



Jack Gallant @gallantlab · May 28



For example, require effect size reports and demand greater evidence for small effects. Require people to report what proportion of individuals show the effect. Separate fit and test sets. Do generalization tests. Test quantitative predictions.



Tal Yarkoni @talyarkoni · May 28



if anything, it should be much *easier* to criticize studies for using NHST when an RR is first submitted for review, than to wait until after the authors are happily trumpeting their $p < .001$ result and can say "but look, it's strong!"



Thomas Yeo @bttyeo · May 28



Replying to @gallantlab @talyarkoni and 4 others

Replacing p values with out of sample prediction will just be shifting the problem from p-hacking to out-of-sample-hacking. My feeling is that many machine learning papers also do not replicate.

- 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 - others

- 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 !

- 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)

Conclusion 1: 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: 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 ?