The replicability problems in Science: It's not the p-values' fault

Yoav Benjamini

Tel Aviv University

NISS Webinar

May 6, 2020

1. The Reproducibility and Replicability Crisis

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

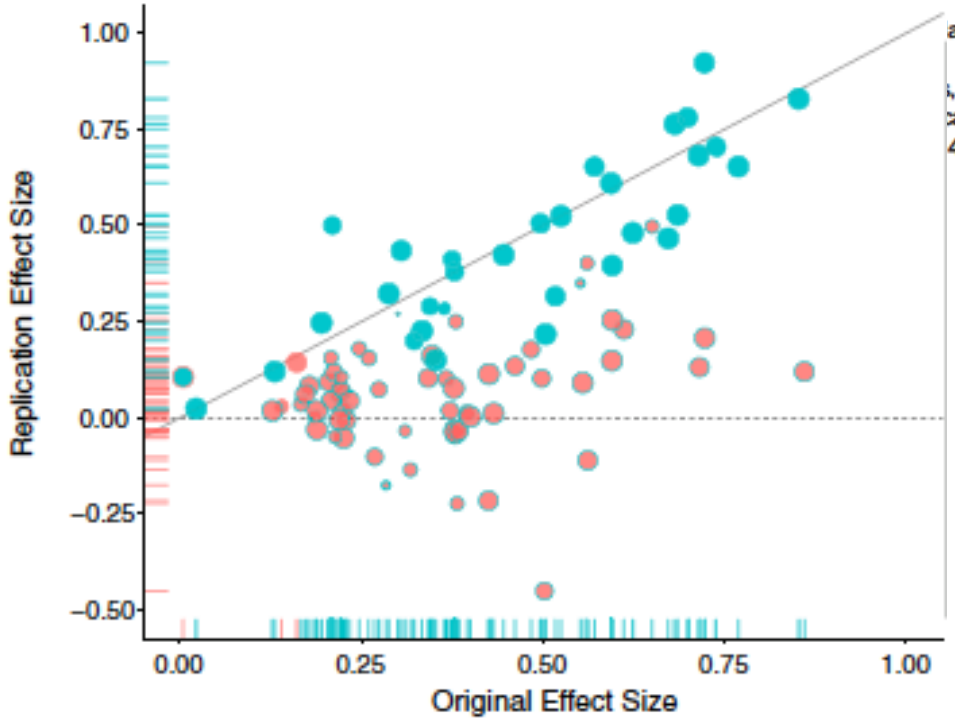


Fig. 3. Original study effect size versus replication effect size (corr
Diagonal line represents replication effect size equal to original effect size.
replication effect size of 0. Points below the dotted line were effects in the o
original. Density plots are separated by significant (blue) and nonsignificant

ational
cally
47%

Open access, freely available online

rch Findings

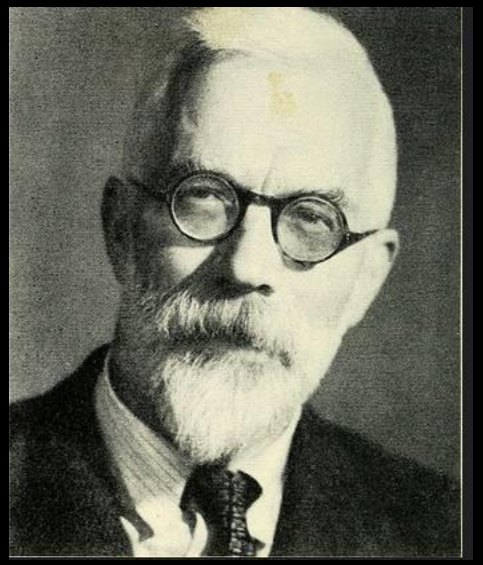
ology Cul

not

weet 1,852

logist, in a

Replicability with significance



"We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results."

Fisher (1935) "The Design of Experiments".

Reproducibility/Replicability

- Reproduce the study: from the original data, through analysis, to get same figures and conclusions
- Replicability of results: replicate the entire study, from enlisting subjects through collecting data, and analyzing the results, in a similar but not necessarily identical way, yet get essentially the same results.

" reproducibilty is the ability to replicate the results…"

in a paper on "reproducibility is not replicability"

We can therefore assure reproducibility of a single study

but only enhance its replicability

Opinion shared by 2019 report of National Academies on R&R

Outline

1. The misguided attack
2. Selective inference: The silent killer of replicability
3. The status of addressing evident selective inference

2. The misguided attack

Psychological Science *"… we have published a tutorial by Cumming ('14), a leader in the new-statistics movement…"*

- 9. Do not trust any p value.
- 10. Whenever possible, avoid using statistical significance or p-values; simply omit any mention of null hypothesis significance testing (NHST).
- 14. …Routinely report 95% CIs…

Editorial by Trafimow & Marks (2015) in Basic and Applied Social Psychology: *From now on, BASP is banning the NHSTP…*

Is it the p-values' fault?

ASA Board's statement about p-values

(Lazar & Wasserstein Am. Stat. 2016):

- Opens: The p-value "can be useful"
- Then comes: a list of "do not" "is not" and "should not" "leads to distortion" – all warnings phrased about the p-value.

Is it the p-values' fault?

It concludes: "In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. "

It is the p-values' fault!

"We're finally starting to get rid of the p-value tyranny"

NEJM editorial July 2019 discussion

When P values are reported for multiple outcomes without adjustment for multiplicity, the probability of declaring a treatment difference when none exists can be much higher than 5%. …

Even when no adjustment for multiplicity is needed,

P values do not represent the probability that the null hypothesis is false… P values provide no information about the variability of an estimated association … P values provide no information about the size of an effect or an association.

What other approaches were mentioned?

Confidence intervals

Prediction intervals

Estimation

Likelihood ratios

Bayesian methods

Bayes factor

Credibility intervals

What other approaches were mentioned in ASA statement?

Confidence intervals

Prediction intervals

Estimation

Likelihood ratios

Bayesian methods

Bayes factor

Credibility intervals

Influenza Vaccination in Pregnancy

JAMA Pediatrics | Original Investigation

Association Between Influenza Infection and Vaccination During Pregnancy and Risk of Autism Spectrum Disorder

(adjusted hazard ratio, 1.20; 95% CI, 1.04-1.39). However, this association could be due to chance ($P = 0.1$) if Bonferroni corrected for the multiplicity of hypotheses tested ($n = 8$). Maternal influenza vaccination in the second or third trimester was not associated with increased ASD risk.

Total No.	1400	443	431	541	195 529
ASD cases, No. (%)	22 (1.57)	8 (1.81)	7 (1.62)	7 (1.29)	3081 (1.60)
Crude hazard ratio (95% CI)	1.00 (0.66-1.53)	1.16 (0.58-2.31)	1.05 (0.50-2.21)	0.81 (0.39-1.70)	1 [Reference]
Adjusted hazard ratio (95% CI)[a]	1.04 (0.68-1.58)	1.18 (0.59-2.37)	1.07 (0.51-2.25)	0.86 (0.41-1.80)	1 [Reference]
Crude risk difference	NE	NE	NE	NE	1 [Reference]
Vaccination					
Total No.	45 231	13 477	17 475	16 095	151 698
ASD cases, No. (%)	765 (1.69)	258 (1.91)	279 (1.60)	260 (1.62)	2338 (1.54)
Crude hazard ratio (95% CI)	1.11 (1.01-1.21)	1.26 (1.10-1.45)	1.03 (0.91-1.18)	1.02 (0.90-1.17)	1 [Reference]
Adjusted hazard ratio (95% CI)[a]	1.10 (1.00-1.21)	1.20 (1.04-1.39)	1.03 (0.90-1.19)	1.03 (0.90-1.20)	1 [Reference]
Crude risk difference	NE	0.40 (0.14-0.63)	NE	NE	1 [Reference]

Abbreviations: ASD, autism spectrum disorder; NE, not estimated.

[a] Hazard ratio adjusted for maternal allergy, asthma, autoimmune conditions, gestational diabetes, hypertension, age, education, race/ethnicity, child conception year, conception season, sex, and gestational age.

Principle 4: Avoid selective reporting of p-values

2. Selective inference

Inference on a selected subset of the parameters that turned out to be of interest **after viewing the data!**

Relevant to all statistical methods – hurting replicability

Out-of-study selection - not evident in the published work

File drawer problem / publication bias

The garden of forking paths, p-hacking, cherry picking

significance chasing, HARKing, Data dredging,

All are widely discussed and addressed

Selective inference

In-study selection - evident in the published work:

Selection by the Abstract

Table

Figure

Selection by highlighting those passing a threshold

$p<.05$, $p<.005$, $p<5*10^{-8}$, *,**,2 fold

Selection by modeling: AIC, C_p, BIC, LASSO,…

In complex research problems - in-study selection is unavoidable!

Approaches for addressing selective inference

A. *Simultaneous over all possible selections*

   FamilyWise Error-Rate

B. *Simultaneous over the selected* *"new"*

C. *Conditional over the selected*

D. *On the average over the selected*

False Discovery Rate & False Coverage Rate

3. The status of addressing evident selective inference

- Clinical trials
  - For drug registration - **Hsien-Ming James Hung Talk**
  - 1st and 2nd stage
  - Old and New NEJM
- Bayesian statistics
- Nature

The status: Clinical trials for drug registration

Phase III trials are analyzed with strict adherence to control the possible effects of selective inference when assessing efficacy

Fuels much statistical research in this area (FWER)

Will be discussed by **Hsien-Ming James Hung today**

What about clinical trials-pre FDA?

Natalizumab, was examined by Ghosh et al (NEJM, 2003) for the treatment of Crohn's disease.

Comparing 3 regimes with placebo; 4 measures of success;

at 5 time points; Total 51 endpoints

1 primary endpoint: Treatment by 2 infusions of 6mg/kg dose
remission measured at week 6

Other 50 described as secondary endpoints

The result for the primary endpoint was not significant ($p= 0.533$);

27 secondary endpoints at $p\leq 0.05$ were considered as discoveries

Study reported as a success

The status: Elsewhere in clinical research?

In depth analysis of 100 papers from the NEJM 2002-2010.

All had multiple endpoints (Cohen and YB '16)

- # of endpoints in a paper 4-167 ; mean=27
- In 80% the issue of multiplicity was entirely ignored: $p \leq 0.05$ threshold (in none fully addressed.)
- All studies designated primary endpoints
- Conclusions based on other endpoints when the primary failed

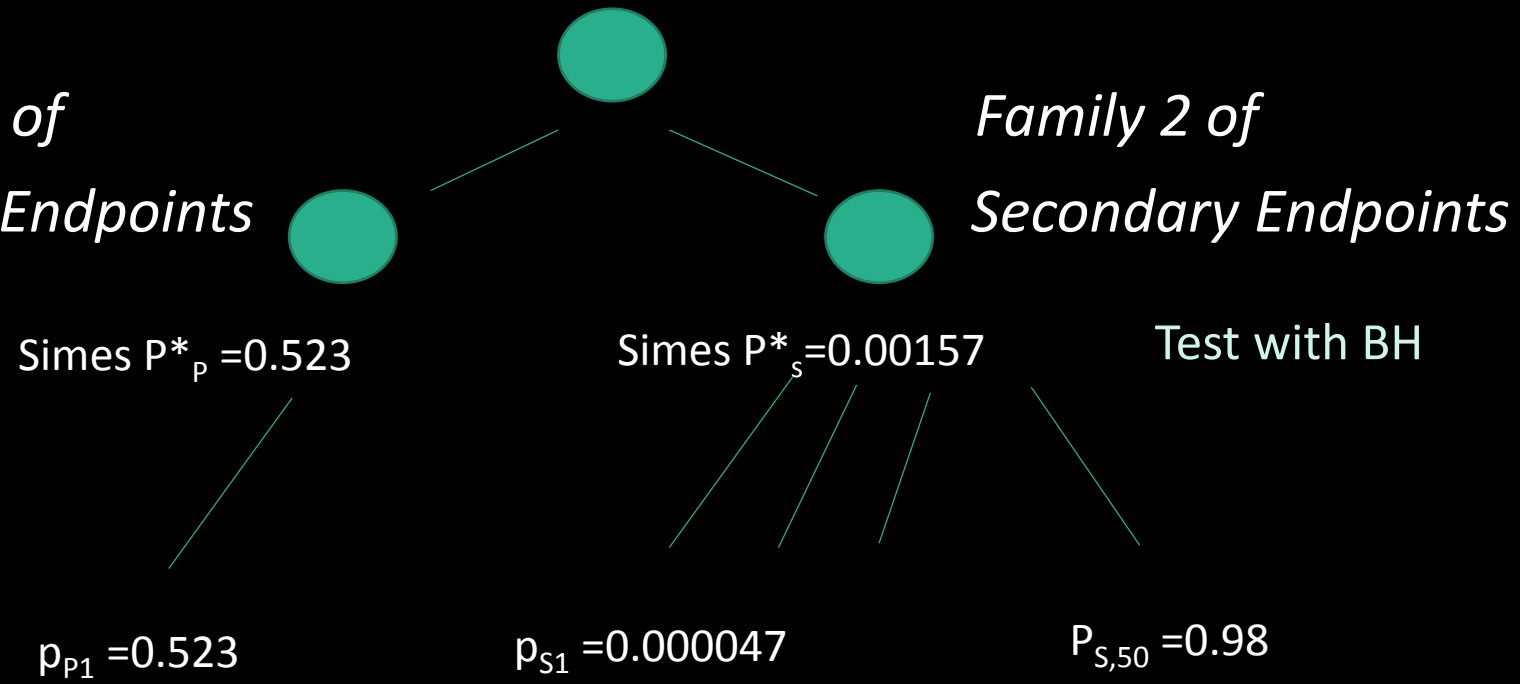
The above reflects most of the published medical research,

Is this why 58% of Phase III trials fail? Nature Reviews

Hierarchical testing of Netalizumab case

Recall one primary endpoint (Family 1),

50 Secondary endpoints (Family 2):



YB & Bogomoov (14)

Hierarchical testing of Netalizumab case

And test with BH the families

Secondary 0.00157 < 1/2*0.05

Primary 0.533 > 0.05

The secondary endpoints tested with BH at 1/2*0.05

12 secondary p-values ≤ 0.05*1/2*12/50 rejected by Hierarchical BH while controlling the error rate (reporting adjusted p-values multiplied by half and FCR intervals.)

Study still a success (YB&Cohen, ’16)

The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D., Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D., Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H

"Some Journal readers may have noticed more parsimonious reporting of P values in our research articles over the past year."

NEJM editorial July 18, 2019

"The new guidelines discuss many aspects of the reporting of studies in the Journal, including a requirement to replace P values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity. "

NEJM editorial (describing Manson et al 2018)

"The n–3 fatty acids did not significantly reduce the rate of either the primary cardiovascular outcome or the cancer outcome. If reported as independent findings, the P values for two of the secondary outcomes would have been less than 0.05;

The Abstract of Manson et al 2018

RESULTS A total of 25,871 participants, including 5106 black participants, underwent randomization. During a median follow-up of 5.3 years, a major cardiovascular event occurred in 386 participants in the n−3 group and in 419 in the placebo group (hazard ratio, 0.92; 95% confidence interval [CI], 0.80 to 1.06; P=0.24). Invasive cancer was diagnosed in 826 participants in the n−3 group and in 797 in the placebo group (hazard ratio, 1.03; 95% CI, 0.93 to 1.13; P=0.56). In the analyses of key secondary end points, the hazard ratios were as follows: for the expanded composite end point of cardiovascular events, 0.93 (95% CI, 0.82 to 1.04); for total myocardial infarction, 0.72 (95% CI, 0.59 to 0.90); for total stroke, 1.04 (95% CI, 0.83 to 1.31); for death from cardiovascular causes, 0.96 (95% CI, 0.76 to 1.21); and for death from cancer (341 deaths from cancer), 0.97 (95% CI, 0.79 to 1.20). In the analysis of death from any cause (978 deaths overall), the hazard ratio was 1.02 (95% CI, 0.90 to 1.15). No excess risks of bleeding or other serious adverse events were observed.

Wu et al, citing results of Manson et al NEJM 2018
Nature Reviews Cardiology (2019)

Fish oil supplementation …

had no significant effect on the composite primary end point of CHD, stroke or death from CVD

but reduced the risk of

total CHD*	(HR 0.83, 95% CI 0.71–0.97),
percutaneous corona intervention	(HR 0.78, 95% CI 0.63–0.95),
total myocardial infarction*	(HR 0.72, 95% CI 0.59–0.90),
fatal myocardial infarction	(HR 0.50, 95% CI 0.26–0.97).

According to the Open Science Framework - Leaders in their efforts to offer tools for pre-registered and transparent research

You should specify

If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?"

You don't have to…

and according to NEJM guidelines you benefit from not

Betinski & Newberger (NEJM 5.12.19) demonstrate that:

All 9 methods that control FWER or FDR control the FWER if all hypotheses are null (even when correlated)

Hence flexibility is not an issue

Harrington's response.

Their assumption requires that the comparisons shown in a manuscript be the only ones examined for possible inclusion.

i.e. If not all are evident, ignore the multiplicity of the evident ones

The post hoc imposition of control of family-wise error rate will reduce the power of the primary test of no overall treatment effect, threatening the original purpose of a study.

Concerned about power?

Control only FDR – but do not ignore!

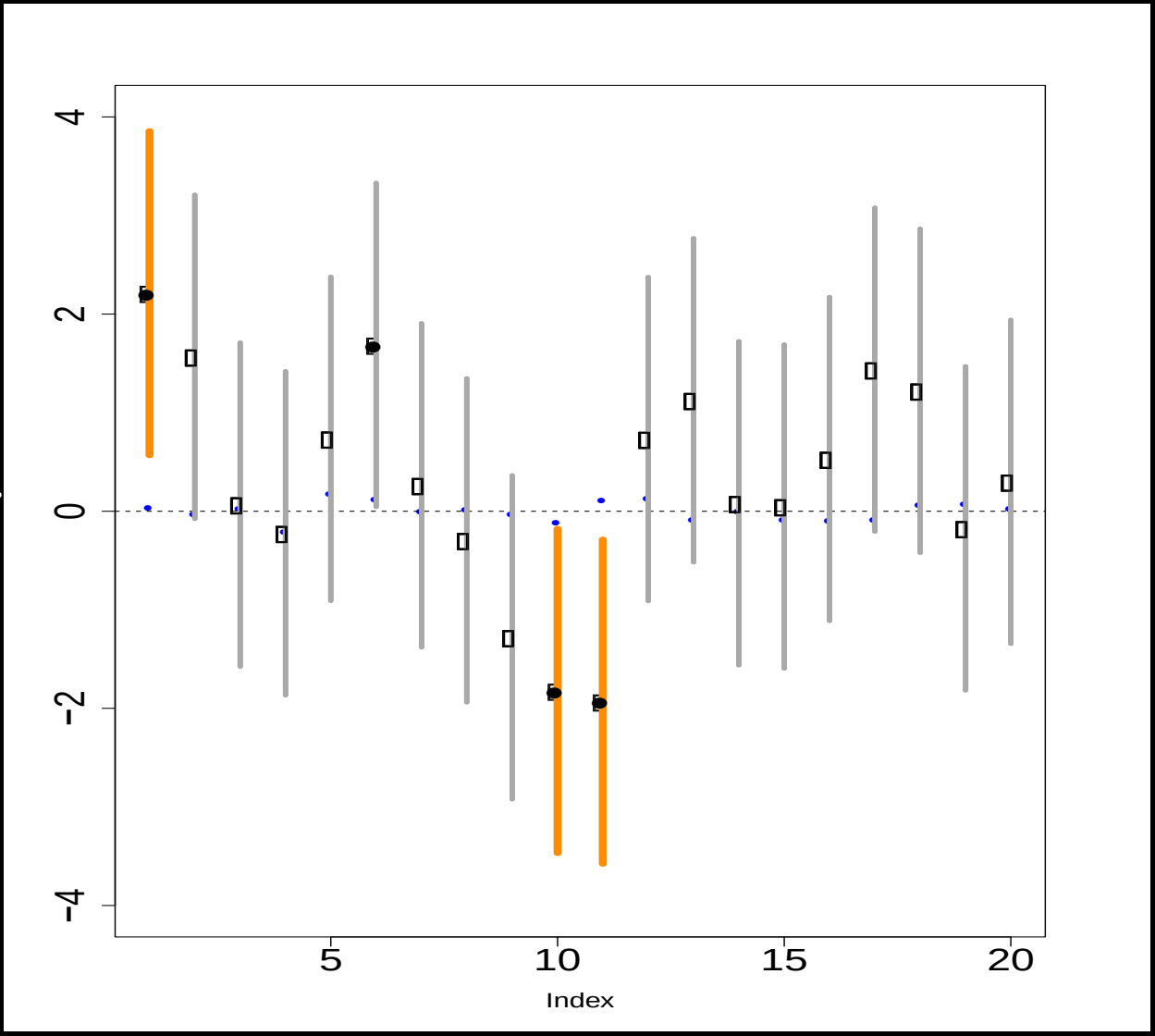
20 parameters to be estimated with 90% CIs

3/20 do not cover

3/4 do not cover when selected

These so selected 4 will tend to fail, or shrink back, when replicated.

FCR CIs have level (1-.1*4/20)100%



Selective inference by CIs is totally ignored

And ASA statement about the p-values is to be blamed

Still usual FWER or FCR CIs have a problem:

Being symmetric

they extend too wide in the direction away from 0

giving a false impression of potential large benefits

The new Simultaneous over the Selected intervals

or the False Cover Rate intervals derived from them,

Can address exactly this concern

Conf. Interval for the largest k–out-of-m

The new *simultaneous* over the *k* so-selected CIs

Extends toward 0 as Bonferroni of *m*

Away from 0 only as Bonferroni of *k*

The control on the average over the selected intervals (FCR) where the above levels are divided by *k*

Towards 0 : usual FCR $(\alpha/2)\,k/m$

Away from 0: regular intervals $\alpha/2$

YB, Hechtlinger, Stark ‘19+

The status: Bayesian statistics

Many ignore: Gelman, Carlin …Westfall Rubin (2013)

Bayesian Data Analysis

Some oppose it Gelman, A., Hill, J., & Yajima, M. (2012).

Why we (usually) don't have to worry

about multiple comparisons.

The underlying theoretical justification

Since we condition on all the data,

any selection after the data is viewed is already reflected in the posterior distribution.

Are Bayesian intervals immune from selection's harms?

Assumed Prior $\mu_i \sim N(0,0.5^2)$; $y_i \sim N(\mu_i,1)$; $i=1,2,\ldots,10^6$ (Gelman's Ex.)

Parameters generated by $N(0,.5^2)$)

Type of 95% confidence/credence intervals	Marginal
Intervals not covering their parameter	5.0%
Intervals not covering 0: Selected	7.3%
Intervals not covering their parameter: Out of the Selected	48%

Not all Bayesians hold this point of view about multiplicity

Connections with FDR in large inferential problems

Genovese & Wasserman, ’02 Storey et al ’03…

Fdr and fdr variations on FDR in empirical Bayes framework

Efron et al ’13 …

Purely Bayes model where selection should be addressed

Yekutieli et al ’13

Thresholding of posterior odds using BH

Scientific Method for the 21st Century: A World Beyond $p < 0.05$

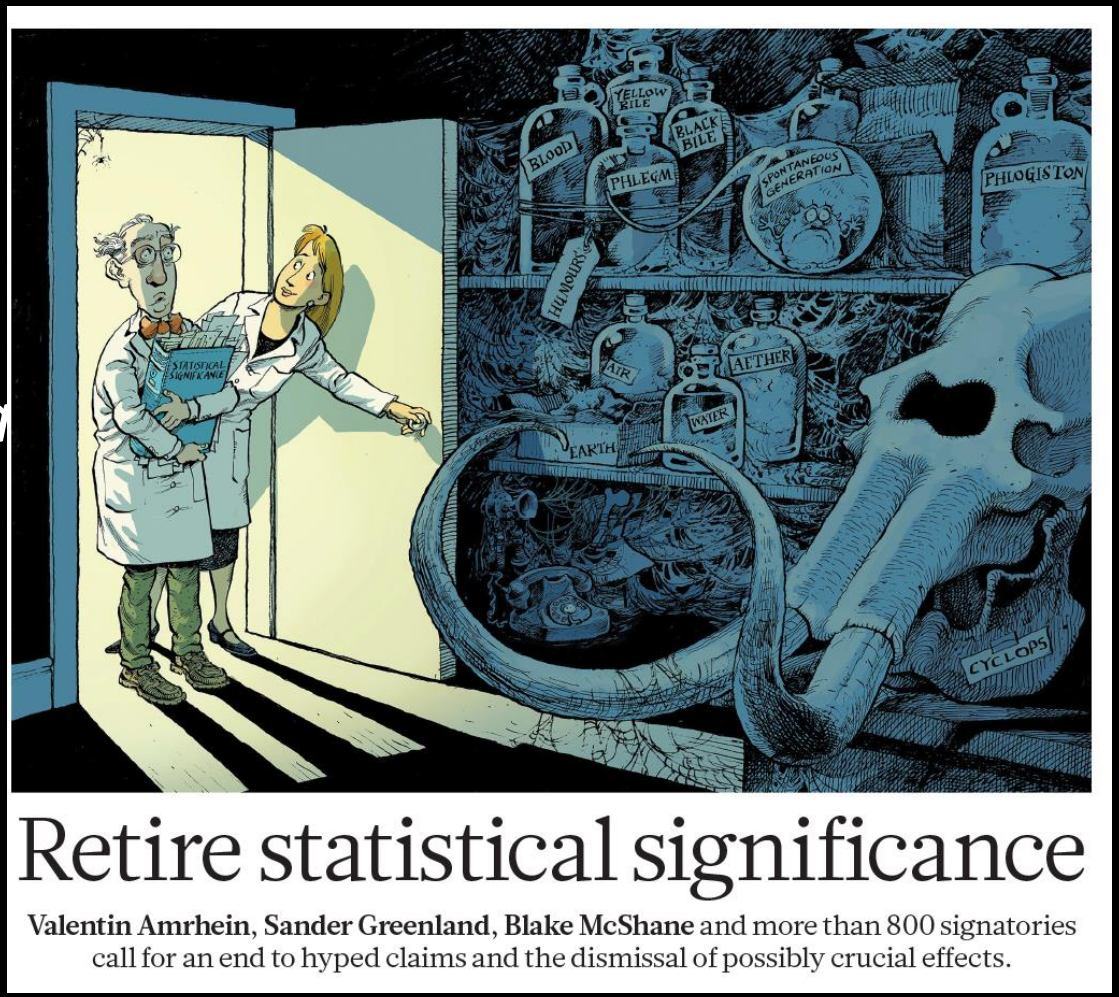
The American Statistician March 2019 Issue

43 papers by participants

An editorial by Ron Wasserstein, Allen Schirm & Nicole Lazar

The Status: Nature Magaszine

'Scientists rise up against statistica significance'



Amrhein, Greenland & McShane ('19)

The Status: Nature Magaszine

But also '~~confidence intervals~~' -> 'plausibility intervals'

- They start with

  "Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a p value is larger than a threshold such as 0.05".

- Continue by objecting to 'Statistical Significance'
- End by objecting to any bright line

Rely on The American Statistician &Hurlbert et al therein

The status of addressing selective inference

Coup de Grâce for a Tough Old Bull:

"Statistically Significant" Expires

Hurlbert, Lavine & Utts object to any bright line

They 'ask': *"how can we address multiple comparisons without a threshold*?"

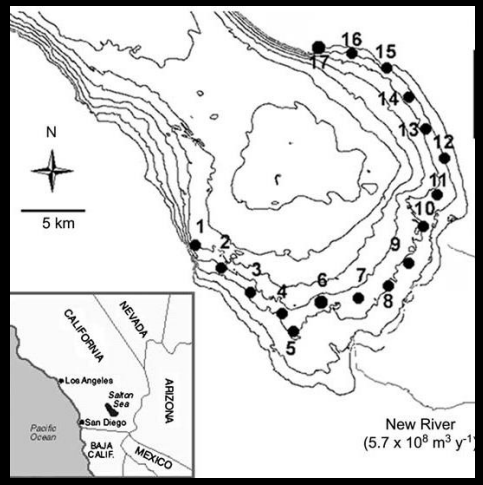
They answer : *"We can't. And should not try*".

Recommend :

"*nuanced reporting*" & "*no need for bright line*" as in Reifel et al '07

The status of addressing selective inference

Influence of river inflows on plankton distribution Around the southern perimeter of the Salton Sea, California



			Per km[a]	R^2	p^b	Per km[a]	R^2	p^b
Dinophyceae								
Gonyaulax grindleyi	45,000	0.12	–5.8	0.60	<0.01	0.69	0.04	0.54
Gyrodinium uncatenum	17,000	0.29	2.8	0.12	0.28	4.2	0.58	0.01
scrippsielloid dinoflagellates	8,300	0.36	4.2	0.24	0.11	2.6	0.24	0.13
Prorocentrum minimum	1,100	0.49	np	np	np	5.2	0.83	<0.01
medium dinoflagellates	970	0.74	4.2	0.48	0.01	1.2	0.11	0.32
tiny dinoflagellates	562	0.89	np	np	np	1.9	0.13	0.28
total dinoflagellate biovolume	–	–	0.23	0.01	0.83	1.9	0.34	0.06
Bacillariophyceae								
Cyclotella sp.	216	0.92	–1.1	0.03	0.58	–1.4	0.51	0.01
Pleurosigma ambrosianum	1,200	1.14	0.69	0.01	0.80	0.93	0.07	0.43
Thalassionema sp.	340	1.62	7.2	0.25	0.10	0.46	0.01	0.77
Cylindrotheca closterium	110	2.48	6.4	0.32	0.06	<0.01	<0.01	0.92
Chaetoceros muelleri	160	3.82	10	0.84	<0.01	np	np	np
total diatom biovolume	–	–	7.4	0.49	0.01	–0.46	0.17	0.22
Raphidophyceae								
Chattonella marina	13,000	0.32	–2.5	0.21	0.13	np	np	np
Cryptophyceae								
cryptomonads	247	1.18	–0.46	0.00	0.90	4.2	0.47	0.02
Euglenophyceae								
Eutreptia sp.	2,400	0.63	–0.23	0.01	0.80	np	np	np
Total phytoplankton	–	–	1.2	0.10	0.33	2.1	0.37	0.05
Chlorophyll *a*	–	–	–0.92	0.04	0.54	4.2	0.66	<0.01
Metazooplankton								
Apocyclops dengizicus	8,630	–	–0.23	0.01	0.81	1.9	0.22	0.15
Balanus amphitrite larvae	19,600	–	14	0.43	0.03	0.93	0.06	0.48
Brachionus rotundiformis	1,130	–	2.3	0.18	0.20	–3.2	0.27	0.13
Neanthes succinea larvae	47,600	–	np	np	np	–0.23	0.002	0.89
Synchaeta spp.	–	–	np	np	np	2.6	0.17	0.21
total zooplankton biovolume	–	–	1.9	0.24	0.13	1.4	0.18	0.19

[a] Calculated as 10^b-1 where $\log A = a + bX$

[b] Significance level of estimated slope (b)

np = not present

Only results with $p \leq 0.1$ Are specifically discussed in the Abstract

Out of 41 results

Ban the use of Abstracts!

Summing up for evident selective inference

Ignoring selective inference evident in the published work is the current status in many branches of science:

Medical Research * Pre-clinical research *Experimental Psychology * Epidemiology * Environmental Research *

And

Among leaders such as NEJM, Open Science Framework, Nature, and even ASA

Sweeping the p-values under the rug worsens the situation

But

Replicability cannot reliably be assessed without

actual replicability efforts by others

Replicating others' work as a way of life

- Every research should have a replicability-check component of a result, considered by the authors important for the research.
- Supported by granting agencies
- The replication effort result is published in short as part of the work
- Meta-analysis of such studies should be simple to perform. Consistency or lack of it, as well as evidence for replicability and generalizability will be assessed

Even independent replication $p<.05$ by 2 investigators is stronger than $p<.005$ and scientifically stronger.

- Special recognition is given to the original authors on whose work replication was attempted

Thanks!

www.replicability.tau.ac.il

1888 1999

The industrialization of the scientific process

1950 2010