The two statistical pillars of replicability: Addressing selective inference and irrelevant variability

Yoav Benjamini

Statistics & O.R., School of Mathematical Sciences
The Sagol School of Neuroscience

Tel Aviv University

The Statistics Wars and Their Casualties
September 2022

Outline

- Tukey’s two pillars of replicability
  - Addressing Selective Inference
  - Relevant Variability
- Some recent evidence
- The relation of StatWars to replicability issues
- Frequentist & Bayesian responses to the pillars
- NEJM guidelines: a casualty of the wars

Tukey’s last published work

A puzzling encyclopedic entry on Multiple Comparisons.

Opens : ``a diversity of issues … that tend to be important, difficult, and often unresolved.”

Details:

- The FDR approach in pairwise comparisons[2]
- The Random Effects vs Fixed Effects analysis [3]

What’s that to do with multiple comparisons?

[1] Jones et al (‘22) International Encyclopedia of Statistics in the Social Sciences

[2] Williams Jones & Tukey (’99) [3] Cornfield & Tukey (’56)

Genes & Behaviour: Crabbe et a (Science, ‘99)

Compared 12 measures across strains at 3 labs

In spite of strict standardization,

Significant Lab*Genotype Interaction

“Thus, experiments characterizing mutants may yield

results that are idiosyncratic to a particular laboratory. ”

We thought that using our computational tools will solve the problem

Comparing 17 measures between 8 inbred strains of mice

At 3 labs: Golani at TAU, Elmer MPRC, Kafkafi NIDA[1]

MCP 2002, [1]Kafkafi et al PNAS 2004

Significance of 8 Strain differences

Behavioral Endpoint	Labs Fixed
Prop. Lingering Time	0.00001
# Progression segments	0.00001
Median Turn Radius (scaled)	0.00001
Time away from wall	0.00001
Distance traveled	0.00001
Acceleration	0.00001
# Excursions	0.00001
Time to half max speed	0.00001
Max speed wall segments	0.00001
Median Turn rate	0.00001
Spatial spread	0.00001
Lingering mean speed	0.00001
Homebase occupancy	0.001
# stops per excursion	0.0028
Stop diversity	0.027
Length of progression segments	0.44
Activity decrease	0.67

Strain x Lab
Interaction
significant

FDR ≤ .05

Strain x Lab
Interaction
not significant

Recalling Mann's warning in "Behavior Genetics in transition"[1] (Science, 94)

"…jumping too soon to discoveries.." (and press discoveries)
"raises the issue of *Replicability"*

Tukey's entry is about replicability of discoveries[1]

Addressing : Selective Inference
The relevant variability

Intensified due to the industrialization of the scientific process

[1] Kafkafi et al (PNAS '04)

Testing the approach[1]

- Took Single lab experimental results involving comparisons between mouse strains from Mouse Phenotyping Database
- Carried similar experiments in 3 labs : JAX, TAUL, and TAUM without standardization
- Used Random Lab Mixed Model Analysis to assess replicability of original result
- Estimated $\gamma^2 = \sigma^2_{GxL} / \sigma^2_{within}$ for each endpoint from Database or from our experiments

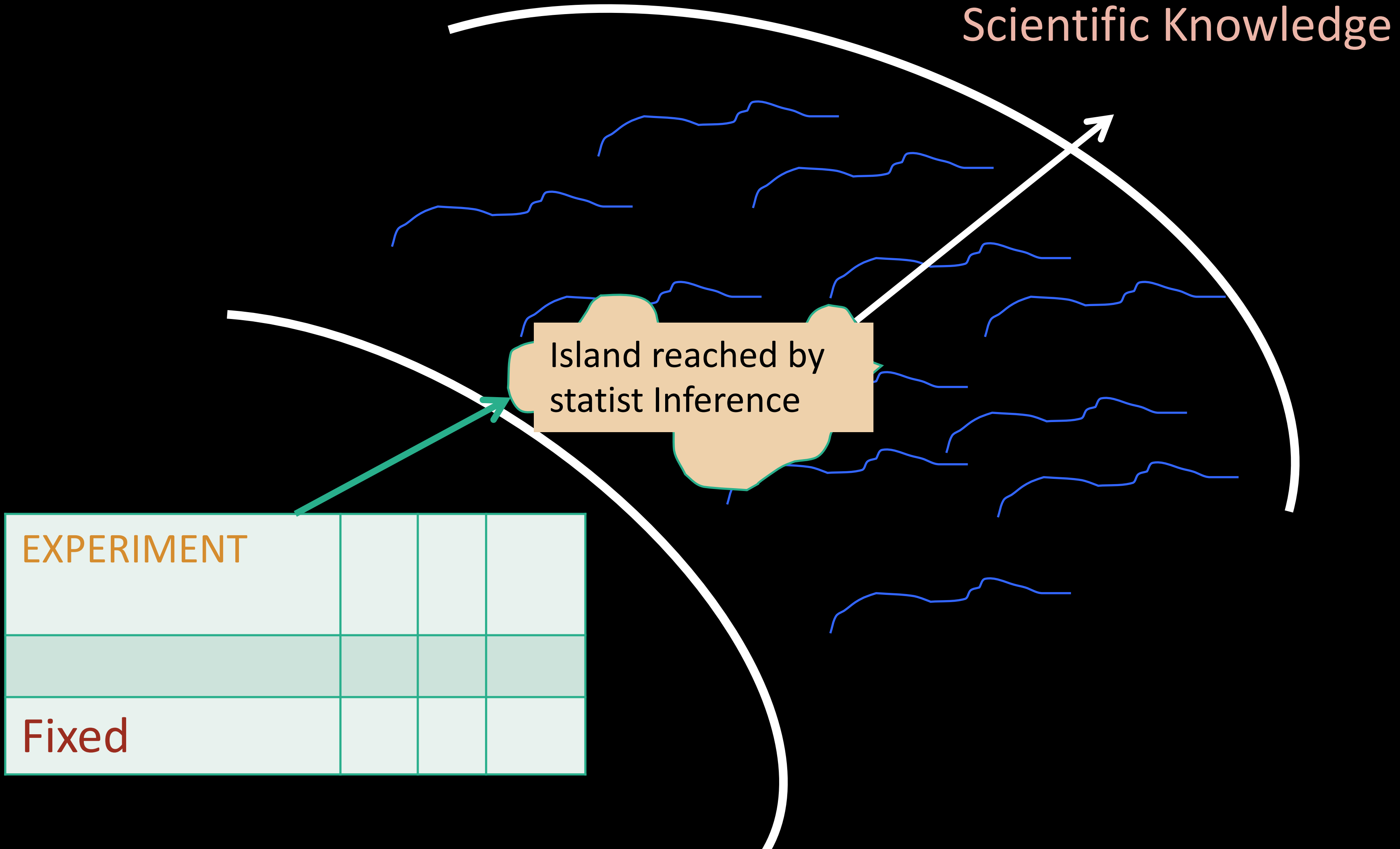
And used it to adjust the single lab results (by inflating sd)

60% of single lab rejected results were non-replicable

12% of adjusted single lab rejected results were non-replicable

Jaljuli, Kafkafi et al ’22+ BioRxiv

Reading ‘56 paper again

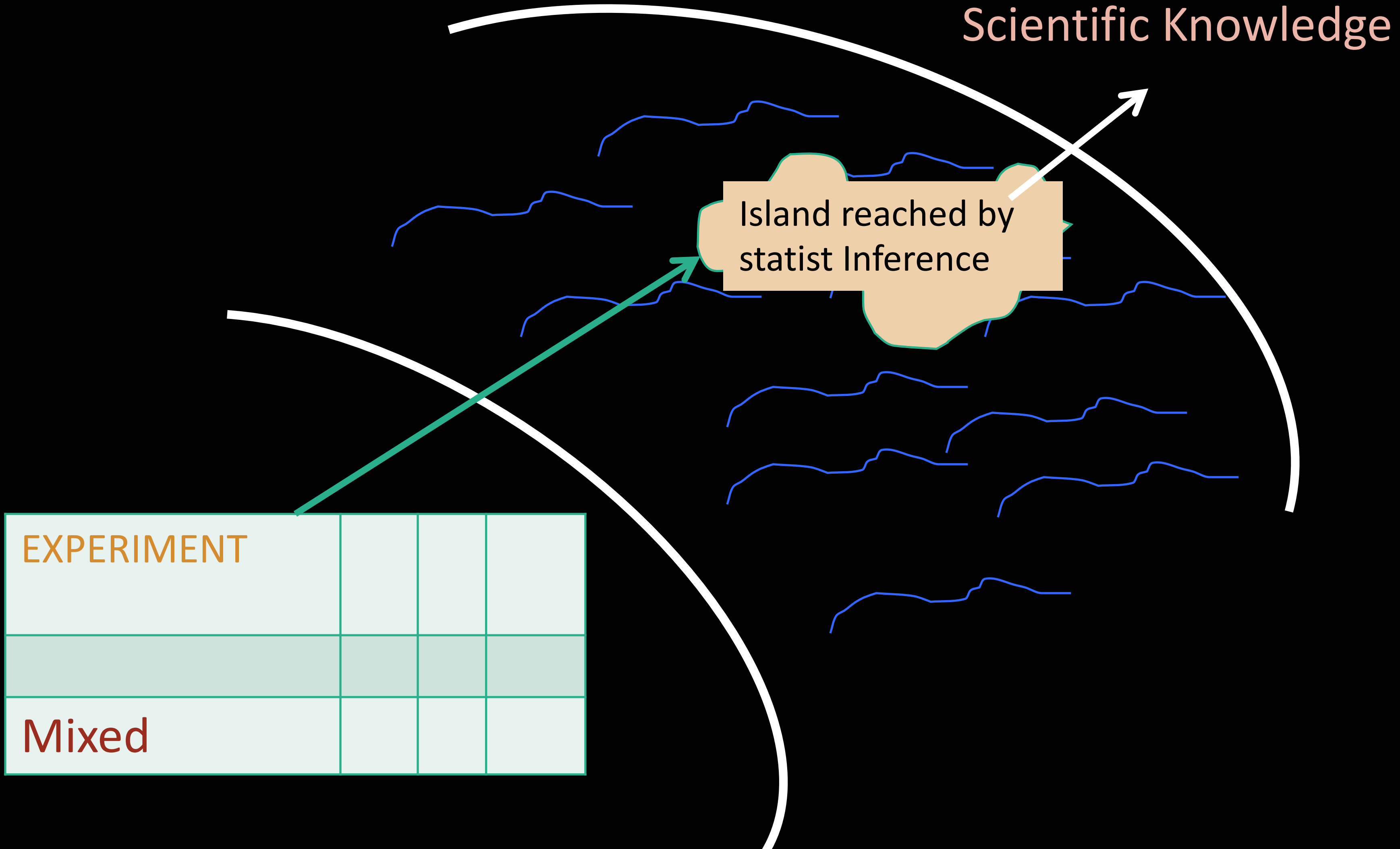


Scientific Knowledge

Island reached by statist Inference

EXPERIMENT			
Fixed			

Reading ‘56 paper again



Scientific Knowledge

Island reached by statist Inference

EXPERIMENT

Mixed

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

e.g. by Transparency & Reproducibility standards

Selective inference

In-study selection - evident in the published work:

Selection by the Abstract, Discussion

Table, Figure

Selection by highlighting those passing a threshold

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

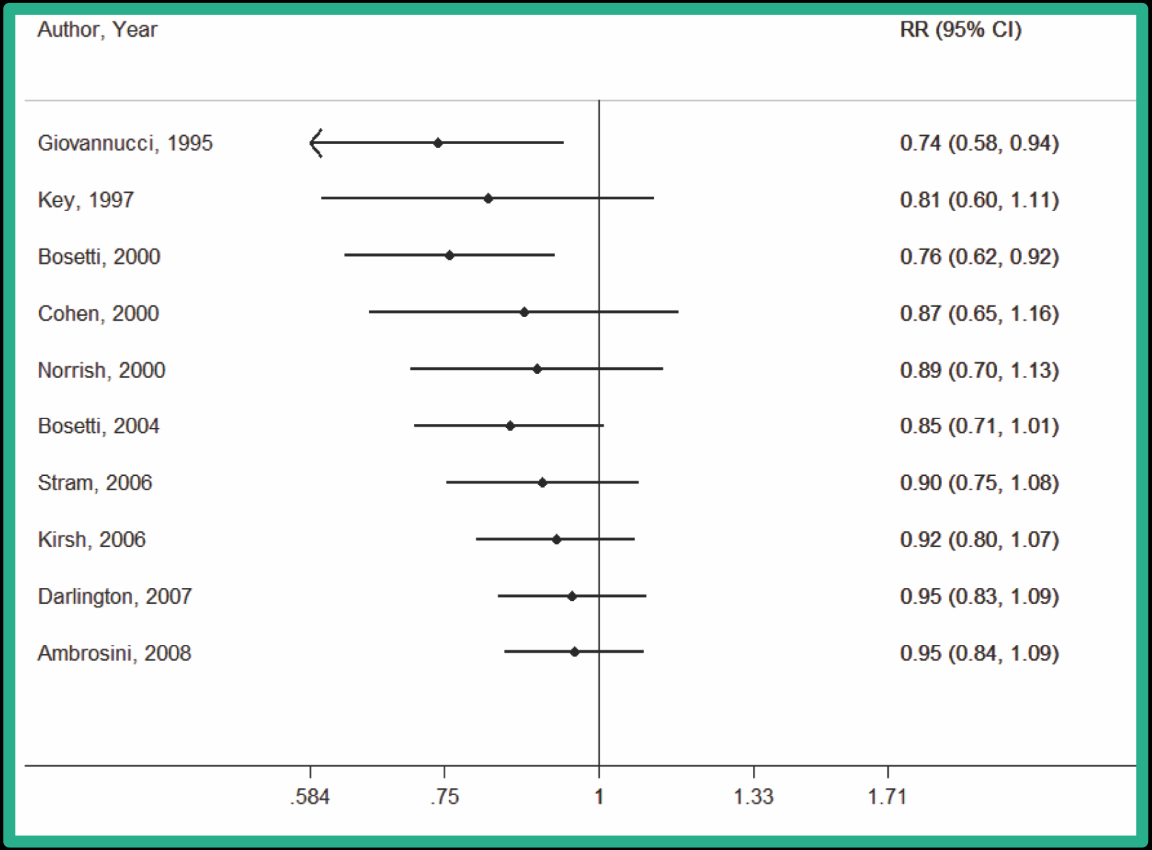
Selection by modeling: AIC, C_p, BIC, LASSO,…

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

Selective inference hampers replicability

- Giovannucci et al. (1995) look for relationships between more than a hundred types of food intakes and the risk of prostate cancer
- The abstract reports three (marginal) 95% confidence intervals (CIs), apparently only for those relative risks whose CIs do not cover 1.

"Eat Ketchup and Pizza and avoid Prostate Cancer"



"Although the pooled RR for raw tomato consumption was initially significant in 1995, this association has remained nonsignificant since 2000 after the addition of 7 studies…" Meta-analysis by Rowles et al (2017)

Any adjustment for selection yields Conf. Intervals covering 1

Error-rates for selective inference

Secondary endpoints from effect of fish oil consumption on CHD, stroke or death from CVD . Manson et al ‘18 used by NEJM editorial

- *Simultaneous over*

  all possible selections

- *Simultaneous*

  over the selected

- *Conditional*

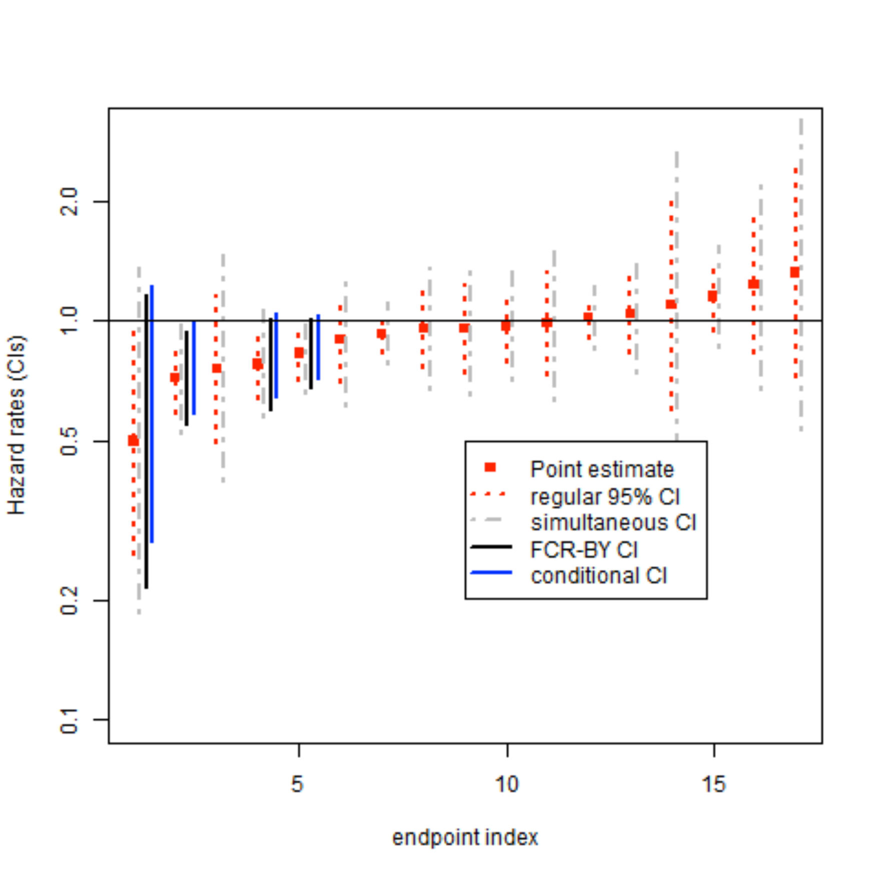
  on being selected

- *On the average*

  over the selected

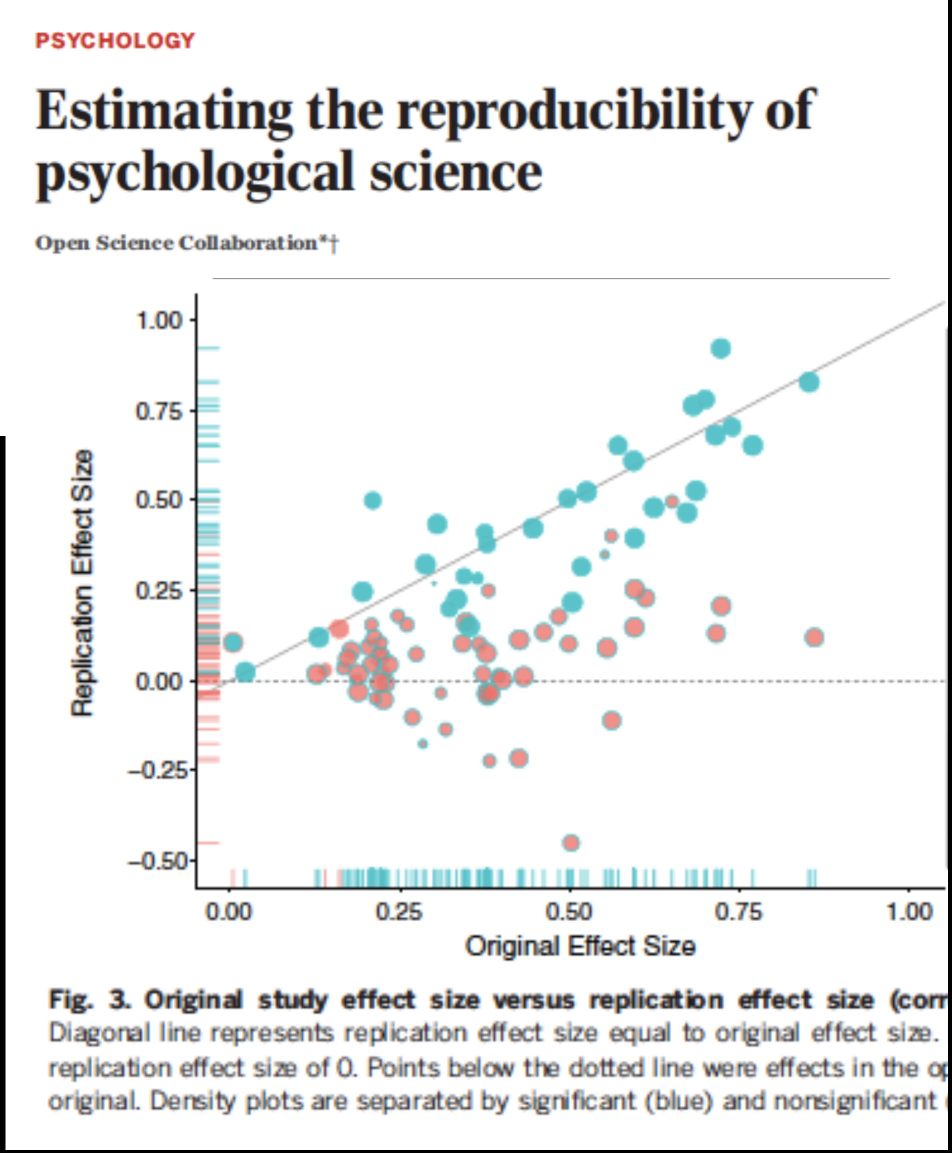
- *On the average*

  over all



Addressing selective inference in psychology

100 replications efforts, 64/100 failed



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 op
original. Density plots are separated by significant (blue) and nonsignificant (

Transparency problems 1/100

Reproducibility problems 6/100

Reproducible & selected p ≤. 05

56/88 (=64%) failed.

Evident selection

Adjusting via hierarchical FDR

22 with $p_{adj} > .05$; and screened

Of them

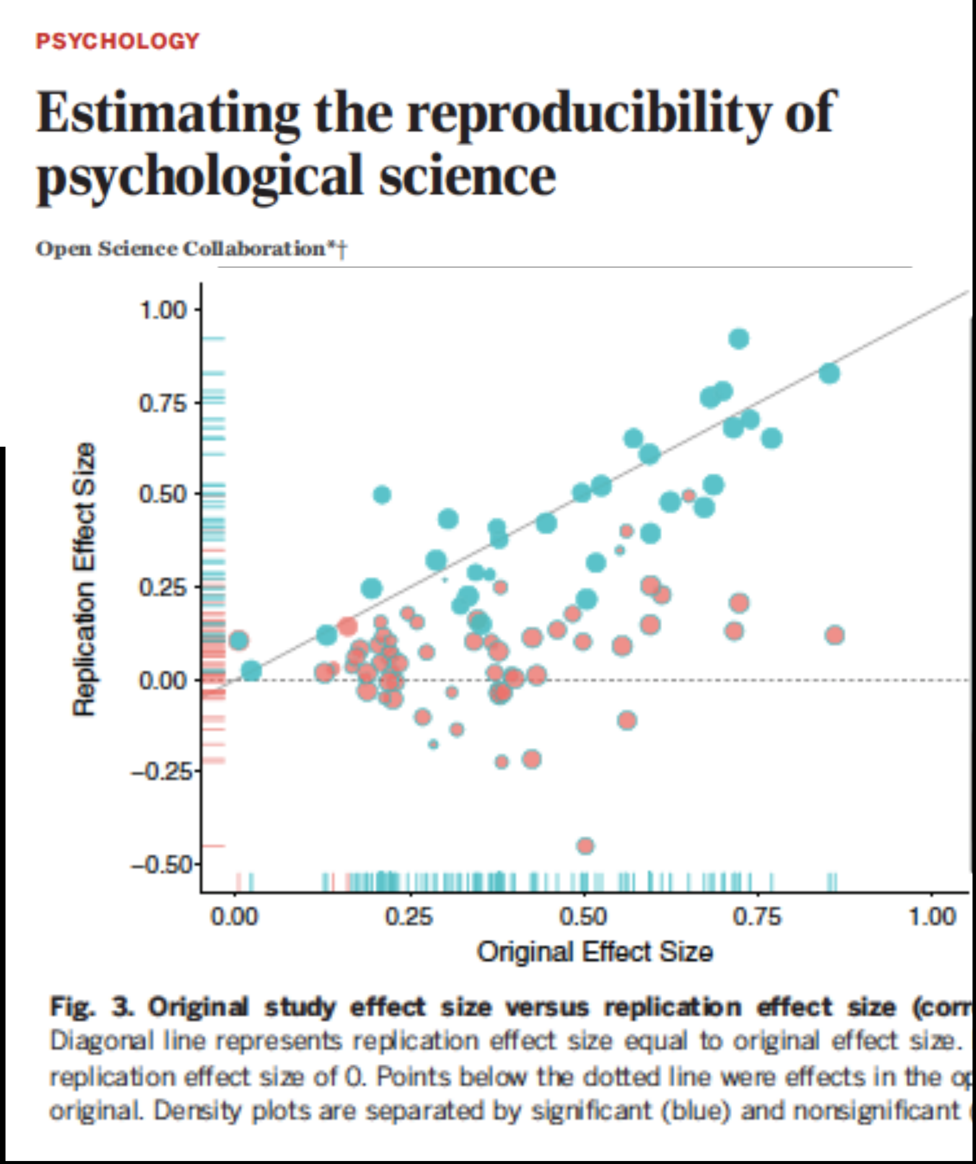
21 non-replicable results

1 replicable discovery lost

Failure rate 36/67 (=52%)

Power loss ~1/31

Addressing selective inference in psychology



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 op original. Density plots are separated by significant (blue) and nonsignificant (

Reducing level to $p \leq 0.005$

Benjamin et al + 200 signatures

32 with $p > .005$; of them

21 non-replicable results

11 replicable discovery

Failure rate 25/47 (=54%)

Power loss =1/3

Zeevi, et al, ('21+)

The statistical wars

A. Identifying problems of non-evident selective inference with the use of p-values and statistical testing

Ending with the ‘New Statistics’ and bans on p-values &NHST

The statistical wars

A. Identifying problems of non-evident selective inference with p-values and statistical testing

B. The 2016 ASA guidelines regarding the p-values[1]

"… some statisticians prefer …to even replace p-values with other approaches"

e.g. Bayes factors, confidence intervals, credence intervals

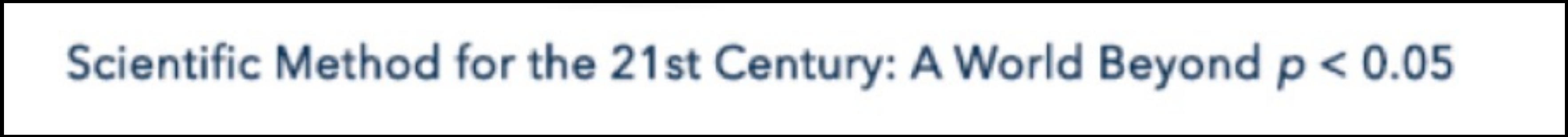
[1] Wasserstein& Lazar (Amm. Stat. '16)

The statistical wars

A. Identifying problems of non-evident selective inference with p-values and statistical testing

B. The 2016 ASA guidelines regarding the p-values[1]

C. The 2019 ASA conference and Editorial[2]



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

Don’t use p<.05; Don’t say “statistically significant”

[1] Wasserstein& Lazar (Am. Stat. ‘16) [2] Wasserstein,Schirm & Lazar (Am. Stat ’19)

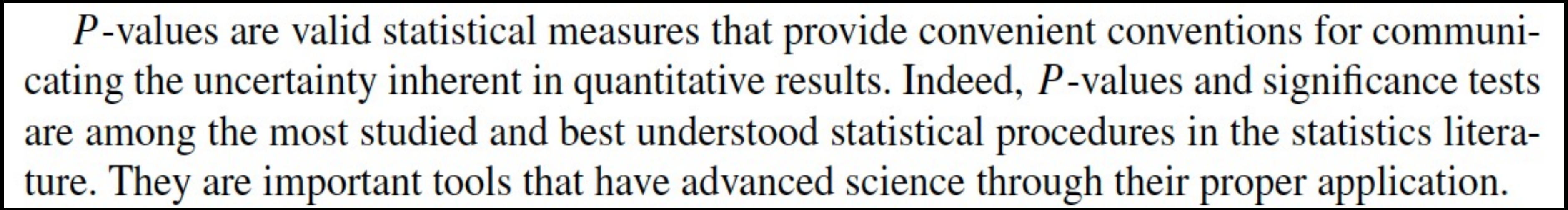
The statistical wars

A. Identifying problems of non-evident selective inference with p-values and statistical testing

B. The 2016 ASA guidelines regarding the p-values

C. The 2019 ASA conference and Editorial

D. The ASA president’s task force statement[3]



P-values are valid statistical measures that provide convenient conventions for communicating the uncertainty inherent in quantitative results. Indeed, *P*-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature. They are important tools that have advanced science through their proper application.

[3] YB et al AOS ‘21

E. The disclaimer

E. By 2022, MacNaughton documented 41 explicit references to 2019 editorial as official ASA policy.

Past-president Kafadar's letter to ASA board required:

- Either the board approves the editorial as policy; or
- A disclaimer is added

The editorial was written by the three editors acting as individuals and reflects their scientific views not an endorsed position of the American Statistical Association.

May 2022 in the online version only

The war goes on

Scientific Reproducibility and Statistical Significance Symposium, Convened by ASA on June 3:

“The goals of the symposium are:”

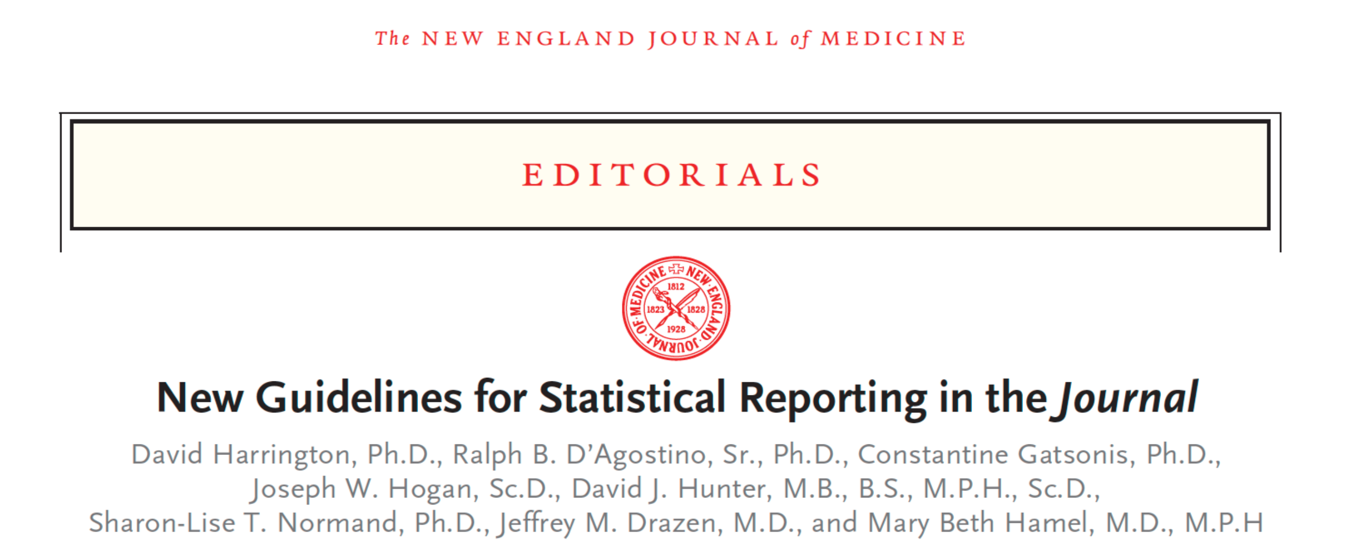
1. To disseminate the stance of the ASA on the appropriate use of results from null hypothesis significance testing

2. To offer alternatives to such testing

3. To discuss changes to publication policies that would benefit both individual scientists and science writ large.

- The war between Bayesian and frequentists was fierce in the 1950’s.
- It receded to co-existence and mutual respect
- The replicability crisis was used by zealous Bayesians to restart the war
- On the eve of the meeting preparing the 2016 statement I expressed my opinion that the statement should be about statistics and replicabiliy in general - not merely focused on the p-value
- Unfortunately, I see no connection between the StatWars and replicability.

The Casualties: NEJM guidelines



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

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%. (July '19)

The Casualties: NEJM guidelines

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

1. *P-values may not be reported (for secondary endpoints) if multiplicity correction method was not specified in the protocol or in the statistical analysis plan*

2. Unadjusted (marginal) 95% CIs reported for all secondary endpoints

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).

The only 4 out of 17 that excluded 1 and were not exploratory.

What’s the problem?

- CI as decision tool

not crossing the no-effect value ⇔ significance testing at 0.05

The issue of multiplicity, as recognized by NEJM, does not disappear

- CIs coverage

Coverage deteriorates: Taking the 17 estimators from above as parameters.

Generated random means with SNR as estimated above times k

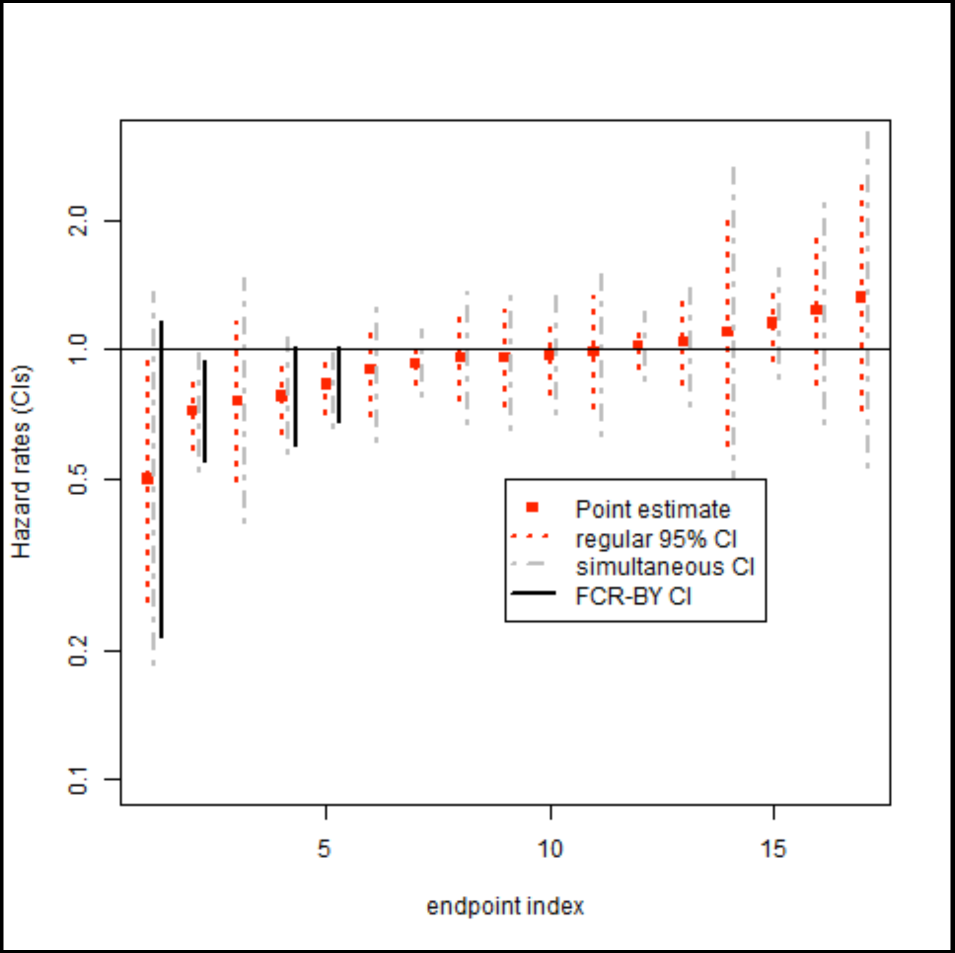
Selected the CIs not crossing above 1 ; 10,000 simulation

Checked the average non-coverage over the so selected

For K=1 0.11 ; For K=0.5 0.18 ; For K=0.01 0.56

Selective inference by CIs is totally ignored

To offer NEJM default guidelines retaining power for exploration



We[1] suggest:

Select if nominal 95%CI does not cross 0 (=log(1))

Assure False Coverage-Rate control over the so selected

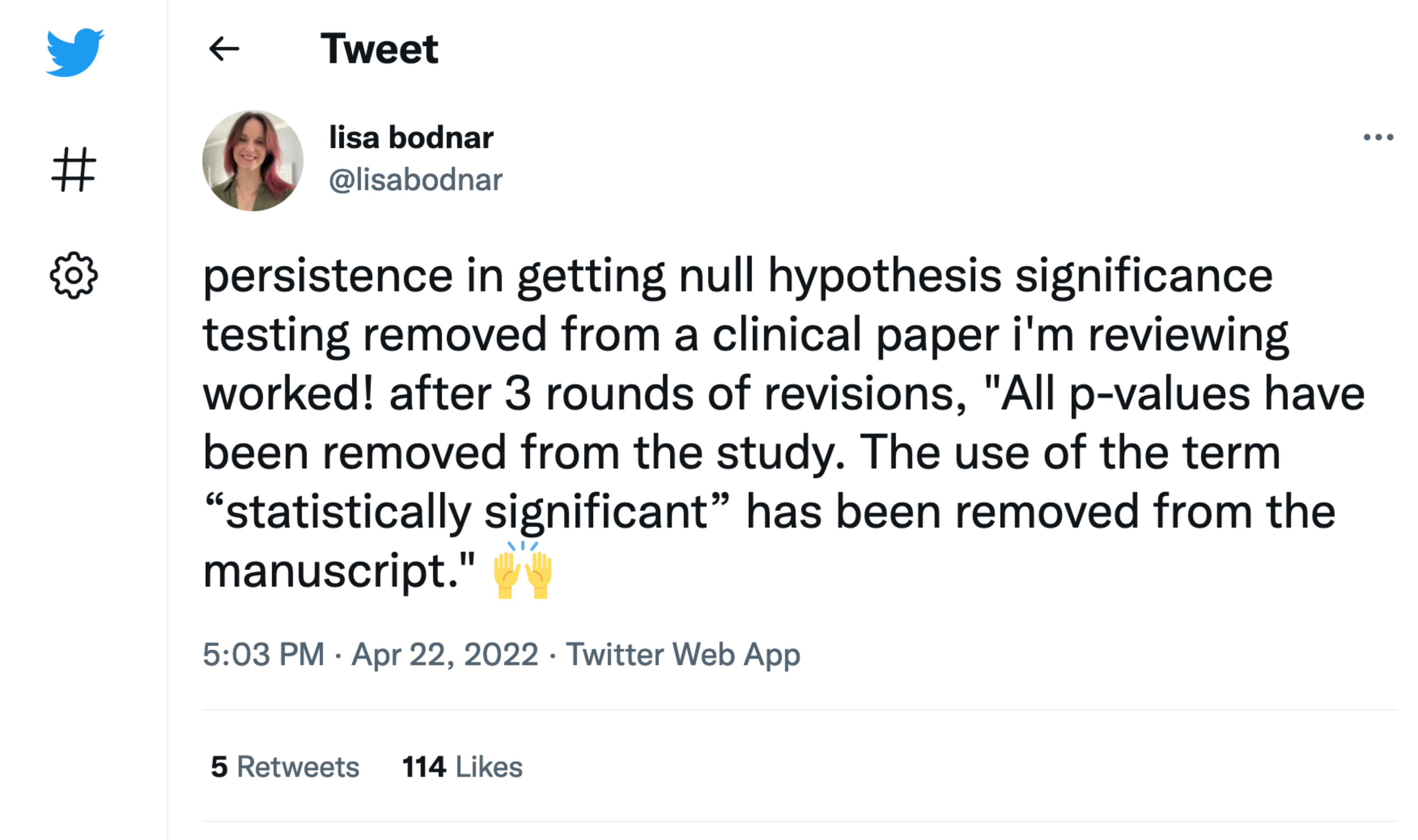
via the general BY-CIs:

* Replace the 4 nominal CIs that do not cross 1 by 95(1-0.05*4/17)% CIs

*For others use nominal CIs

[1] YB, Heller, Panagiotou arXiv '21

The Casualties: 2



Tweet

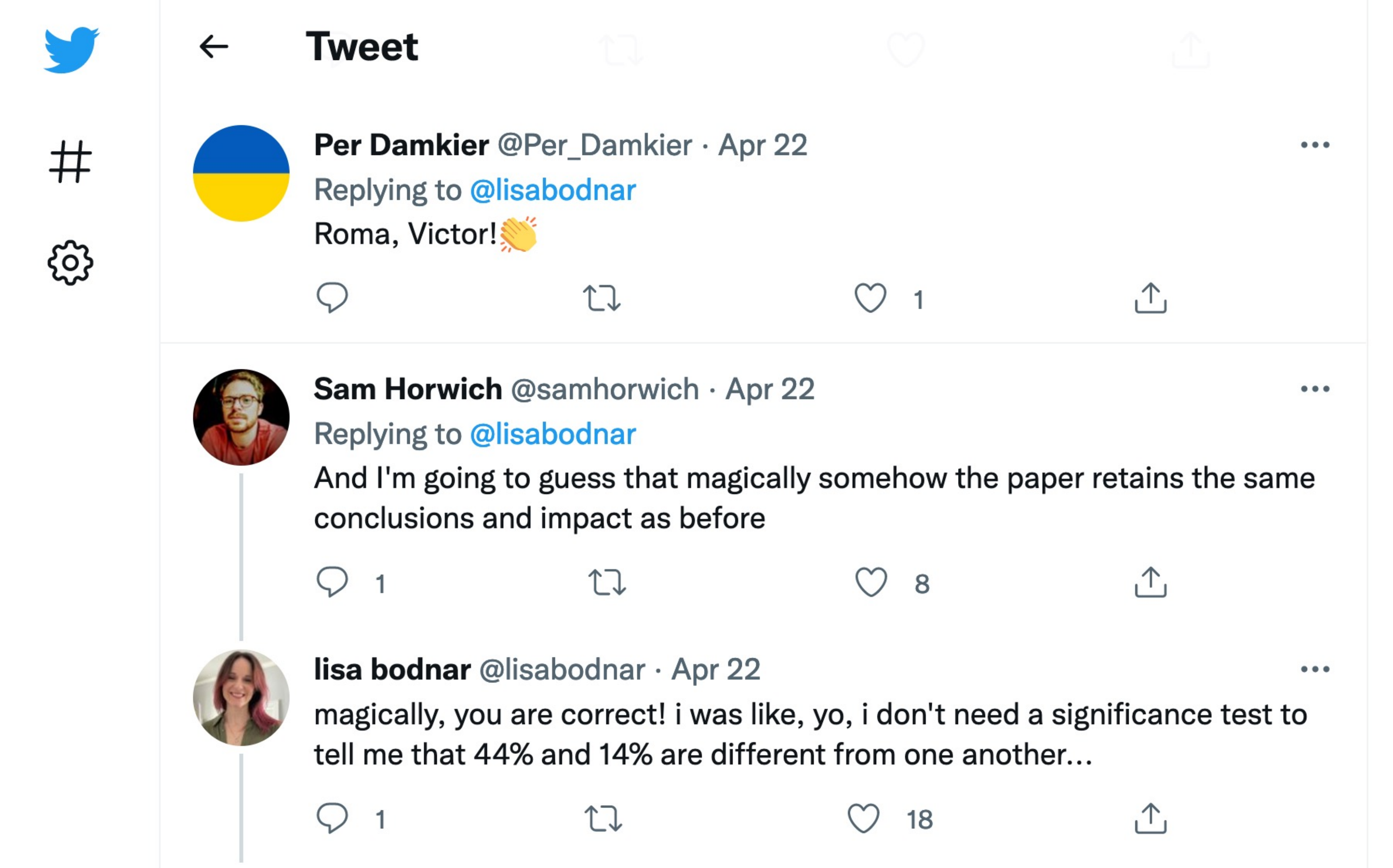
lisa bodnar
@lisabodnar

persistence in getting null hypothesis significance testing removed from a clinical paper i'm reviewing worked! after 3 rounds of revisions, "All p-values have been removed from the study. The use of the term “statistically significant” has been removed from the manuscript." 🙌

5:03 PM · Apr 22, 2022 · Twitter Web App

5 Retweets **114** Likes

An epidemiologist with 13k followers



Per Damkier @Per_Damkier · Apr 22

Replying to @lisabodnar

Roma, Victor!👏

Sam Horwich @samhorwich · Apr 22

Replying to @lisabodnar

And I'm going to guess that magically somehow the paper retains the same conclusions and impact as before

lisa bodnar @lisabodnar · Apr 22

magically, you are correct! i was like, yo, i don't need a significance test to tell me that 44% and 14% are different from one another...

From a Lancet reviewer

- How can statistical hypotheses and strategies for addressing multiplicity issues be pre-specified in an observational study, which typically requires exploratory analyses to reduce bias?
- Can an author claim to have performed a confirmatory test without this pre-specification?

Logical Fallacy:

Confirmatory analysis => Multiplicity adjustment

Therefore:

Multiplicity adjustment => Confirmatory analysis

The 2 pillars in Frequentist & Bayesian eyes

Addressing the relevant variability

Frequentists Ignored by many but increasing awareness

Bayesians Recognized (Hierarchical modelling)

Addressing evident selective inference

Frequentists Recognized more for p-values less for CIs & exploratory research

Bayesian Ignored as a matter of principle

E.g. Gelman, Hill & Yajima, M. (‘12)

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%

YB HDSR '19

From Gelman’s Blog August ‘22

Eric van Zwet offers a solution to the above example:

“We can mix the N(0,0.5) with almost any wider normal distribution with almost any probability and then very large effects will hardly be shrunken.

He demonstrates it by the prior 0.99*N($0,0.5^2$)+0.01*N($0,6^2$)

Of 741 credible intervals not covering 0, the proportion not covering the parameter is 0.07 (CI: 0.05 to 0.09)

Gelman response: I continue to think that Bayesian inference completely solves the multiple comparisons problem.

So, I generated data from this prior to which I add 5% N(4.05^2)

And got 29% of the so selected not covering their parameter. (with an order of magnitude power loss relative to frequentist FDR adjusted testing and FCR CIs)

My point: Bayesians should worry about selective inference if they care about replicability,

And cannot hide behind the theoretical guarantees.

Some Bayesians do that

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

Take away massages

Replicability can be enhanced mainly by addressing

Selective inference

- Evident selective inference is as harmful as non-evident
- Needed in exploratory research even when spool is mall
- Needed for CIs
- ASA attitude against p-values and statistical significance is political and harms replicability

The relevant variability

- Prefer random effect (mixed model) analysis
- Many small studies are better than one/few large ones

Thanks to JWT for the insight

Take away massages

Most Bayesians ignore selective inference

But appreciate addressing the relevant variability

For frequentists it's the opposite

In both research communities practitioners try to avoid addressing them,

Until the research complexity is so large that selective inference is addressed

Until results are so non-replicable that the relevant variability is addressed

This usually takes too long

Thanks to JWT for the insight

Thanks!

www.replicability.tau.ac.il

1888 1999

The industrialization of the scientific process

1950 2010