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

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 opposite direction.
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.”

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”
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
### Principle 4: Avoid selective reporting of p-values

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

<table>
<thead>
<tr>
<th></th>
<th>1400</th>
<th>443</th>
<th>431</th>
<th>541</th>
<th>195 529</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD cases, No. (%)</td>
<td>22 (1.57)</td>
<td>8 (1.81)</td>
<td>7 (1.62)</td>
<td>7 (1.29)</td>
<td>3081 (1.60)</td>
</tr>
<tr>
<td>Crude hazard ratio (95% CI)</td>
<td>1.00 (0.66-1.53)</td>
<td>1.16 (0.58-2.31)</td>
<td>1.05 (0.50-2.21)</td>
<td>0.81 (0.39-1.70)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Adjusted hazard ratio (95% CI)</td>
<td>1.04 (0.68-1.58)</td>
<td>1.18 (0.59-2.37)</td>
<td>1.07 (0.51-2.25)</td>
<td>0.86 (0.41-1.80)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Crude risk difference</td>
<td>NE</td>
<td>NE</td>
<td>NE</td>
<td>NE</td>
<td>1 [Reference]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>45 231</th>
<th>13 477</th>
<th>17 475</th>
<th>16 005</th>
<th>151 698</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD cases, No. (%)</td>
<td>765 (1.60)</td>
<td>258 (1.91)</td>
<td>279 (1.60)</td>
<td>260 (1.62)</td>
<td>2338 (1.54)</td>
</tr>
<tr>
<td>Crude hazard ratio (95% CI)</td>
<td>1.11 (1.01-1.21)</td>
<td>1.26 (1.10-1.45)</td>
<td>1.03 (0.91-1.18)</td>
<td>1.02 (0.90-1.17)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Adjusted hazard ratio (95% CI)</td>
<td>1.10 (1.00-1.21)</td>
<td><strong>1.20 (1.04-1.39)</strong></td>
<td>1.03 (0.90-1.19)</td>
<td>1.03 (0.90-1.20)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Crude risk difference</td>
<td>NE</td>
<td><strong>0.40 (0.14-0.63)</strong></td>
<td>NE</td>
<td>NE</td>
<td>1 [Reference]</td>
</tr>
</tbody>
</table>

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

*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.
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

\[ p < 0.05, \; p < 0.005, \; p < 5 \times 10^{-8}, \; *, **, 2 \text{ fold} \]

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):

**Family 1 of Primary Endpoints**

- Simes $P^*_p = 0.523$
- $p_{p1} = 0.523$

**Family 2 of Secondary Endpoints**

- Simes $P^*_s = 0.00157$
- $p_{s1} = 0.000047$

Test with BH

$p_{s,50} = 0.98$

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)
Some Journal readers may have noticed more parsimonious reporting of P values in our research articles over the past year.”
“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.”
“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;
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 620 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 endpoint 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 $(1-1\times 4/20)100\%$
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


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 $\mu_i \sim \text{N}(0, 0.5^2); \quad y_i \sim \text{N}(\mu_i, 1); \quad i=1,2,\ldots,10^6$ (Gelman’s Ex.)

Parameters generated by $\text{N}(0, .5^2)$

<table>
<thead>
<tr>
<th>Type of 95% confidence/credence intervals</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervals not covering their parameter</td>
<td>5.0%</td>
</tr>
<tr>
<td>Intervals not covering 0: <strong>Selected</strong></td>
<td>7.3%</td>
</tr>
<tr>
<td>Intervals not covering their parameter: <strong>Out of the Selected</strong></td>
<td>48%</td>
</tr>
</tbody>
</table>
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
The American Statistician March 2019 Issue

43 papers by participants
An editorial by Ron Wasserstein, Allen Schirm & Nicole Lazar
'Scientists rise up against statistical significance'

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Amrhein, Greenland & McShane ('19)
The Status: Nature Magazine

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

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
The industrialization of the scientific process

1888  1999

1950  2010

Thanks!

www.replicability.tau.ac.il