Inference after selection with the False Discovery Rate

Yoav Benjamini

Tel Aviv University

CERN, February 2020

Outline

The the replicability crisis and selective inference

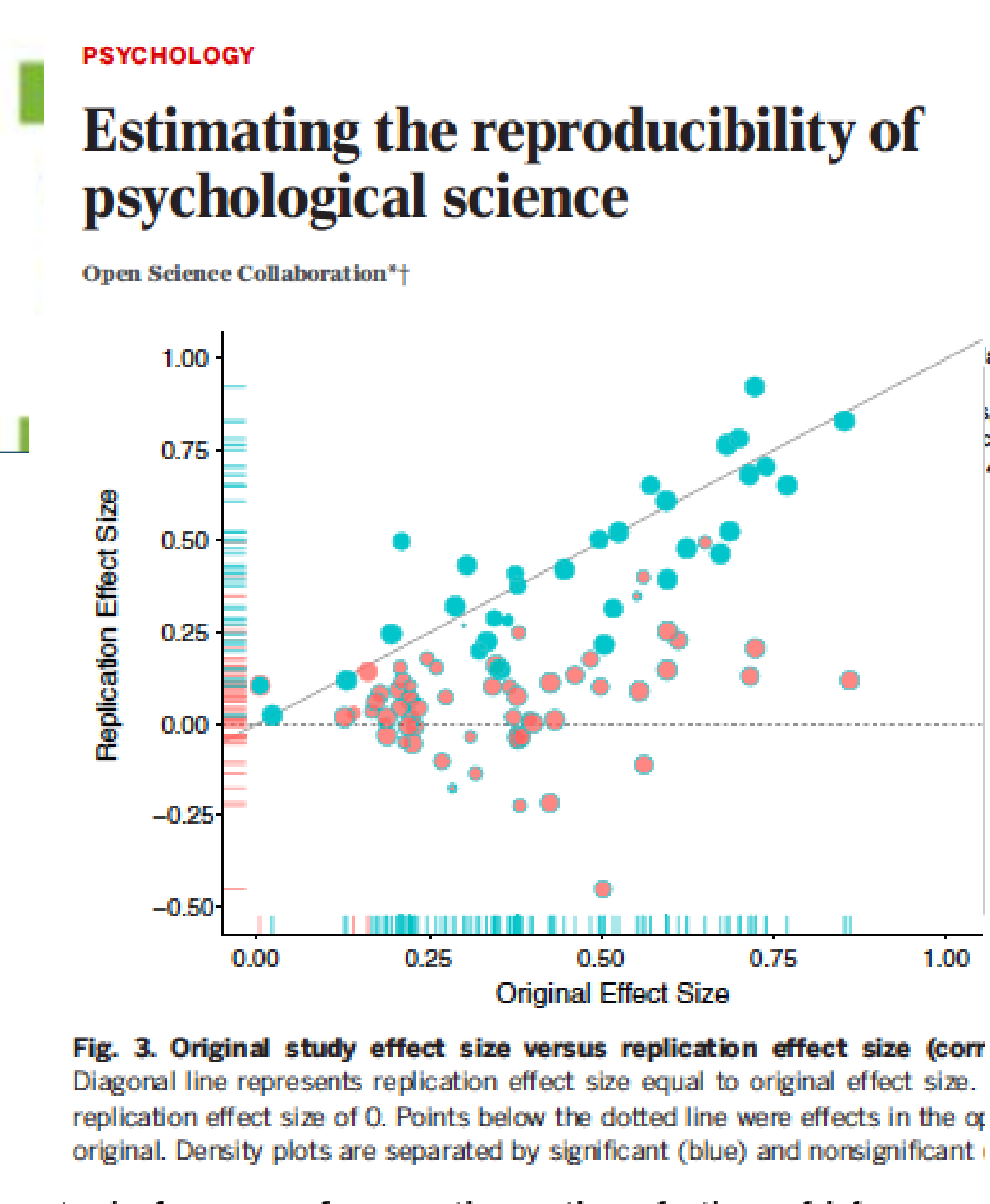
Addressing selective inference by

- Simultaneous inference
- False Discovery rate/ On the average over the selected
- Conditional inference

Selective inference on a database

Final comments on the use of statistics in science

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

ational

ally

47%

Open access, freely available online

rch Findings

iology Cul

not

weet 1,852

logist, in a

known as

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.

(Biostatistics, Editorial 2010, Nature Editorial 2013, NSA 2019)

“ 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

Enhancing Replicability

At the level of the single study?

1. Well and transparently designed experiment
2. Reproducible data analysis and computation

(Nature ’13, NIH in Nature ’14, Science ’14)

All agree that there is need for

3. Statistical methodology that enhances replicability

But what is it?

What problems should it address?

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. Prefer 95% CIs to SE bars. Routinely report 95% CIs…

Basic and Applied Social Psychology (Trafimow & Marks '15)

From now on, BASP is banning the NHSTP

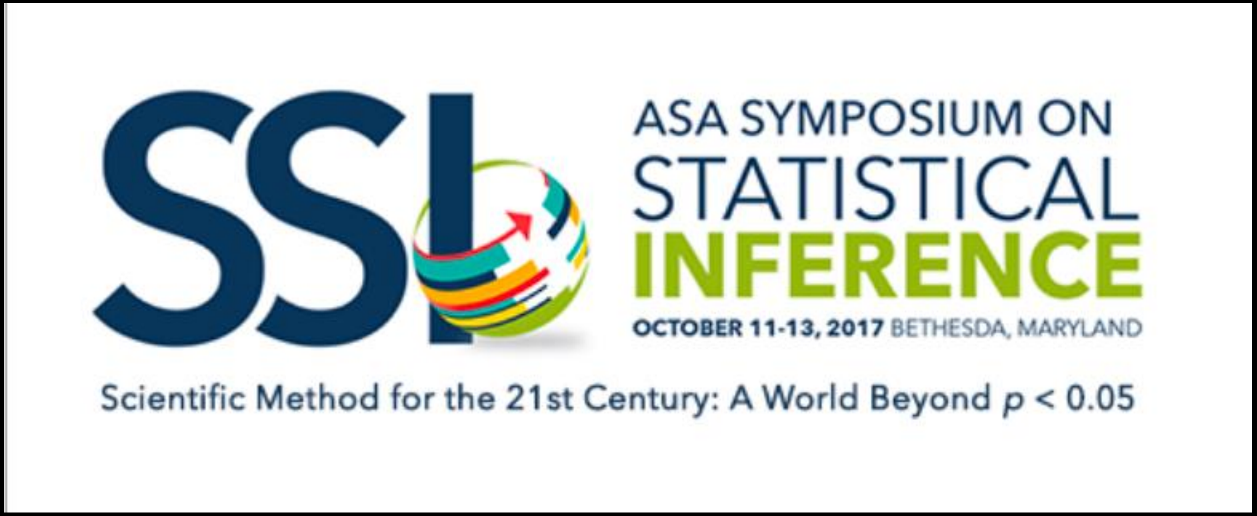
American Statistical Association Statement about the p-value ('16)

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.

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!



SSI ASA SYMPOSIUM ON STATISTICAL INFERENCE

OCTOBER 11-13, 2017 BETHESDA, MARYLAND

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

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

The American Statistician March 2019 Issue

43 papers by participants

A personal editorial by authors of the p-value statement

What other approaches were mentioned?

Confidence intervals

Prediction intervals

Estimation

Likelihood ratios

Bayesian methods

Bayes factor

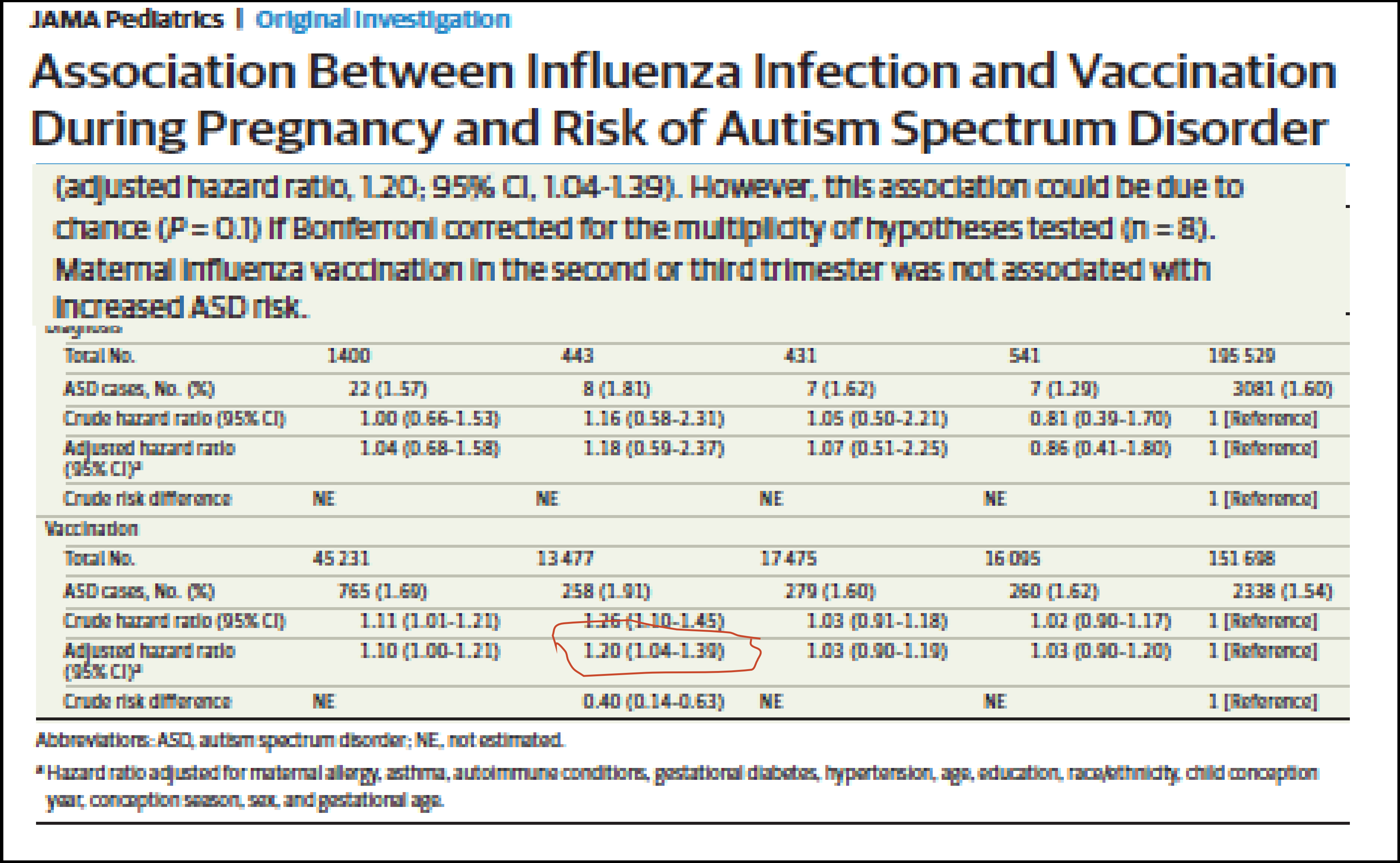
Credibility intervals

Epidemiology: a p-values free zone

- 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"

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.

Diagnosis					
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

Two main statistical challenges to replicability

which are relevant to all statistical methods

A. Addressing selective inference

LEE-like

B. Addressing the relevant variability

Unknown-Systematics

3. Inference on the selected

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

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

File drawer problem / publication bias

The garden of forking paths, p-hacking,

significance chasing, HARKing, Data dredging,

Widely discussed and addressed by Transparency of data, analysis, software: Open & Reproducible Research

In-study selection - evident in the published work:

Selection by the Abstract

Table

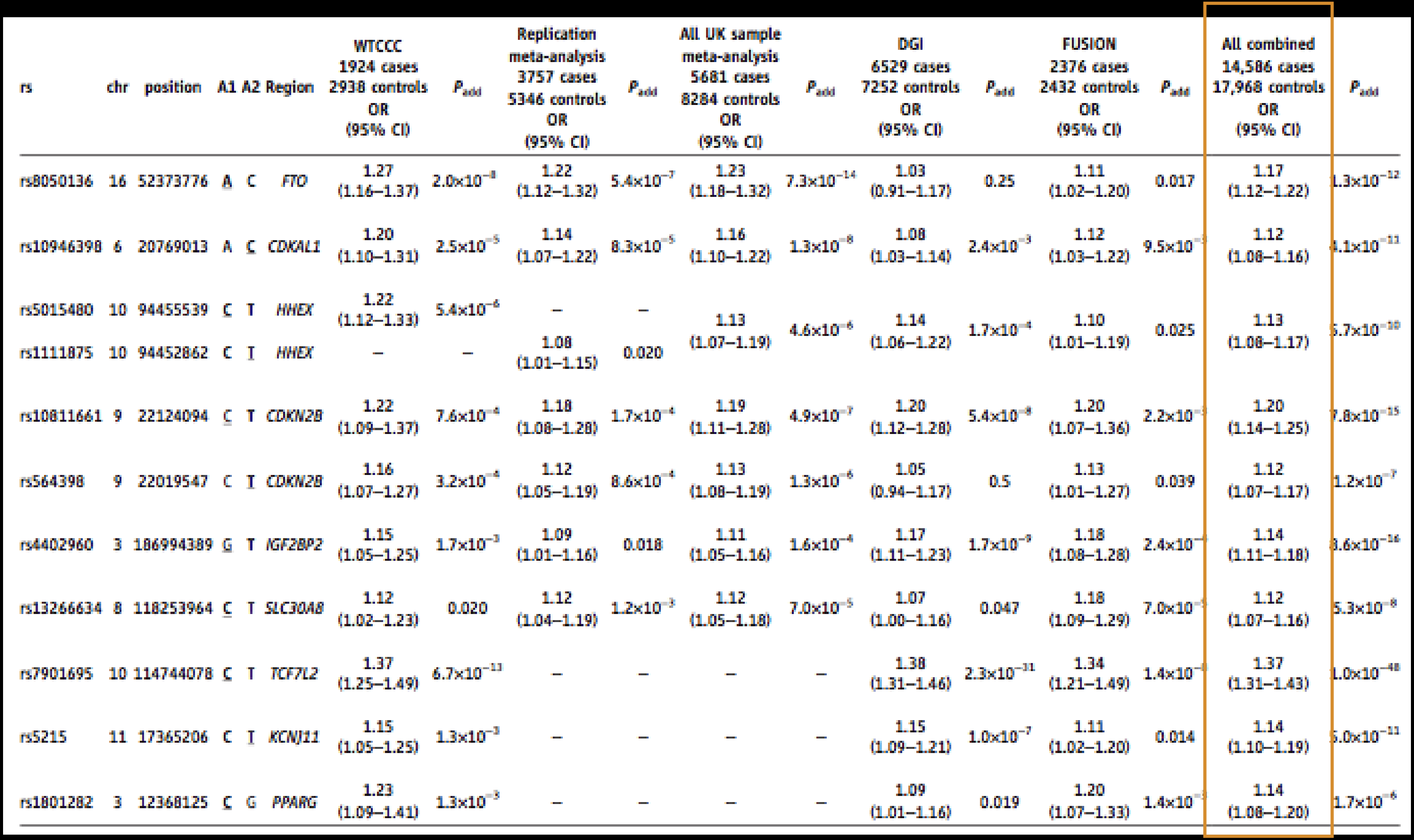
Figure

Selection by highlighting those passing a threshold

$p<.05$, $p<.005$, $p<5*10^{-8}$, $3*10^{-7}$,*,**,2 fold

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

Selection by a Table



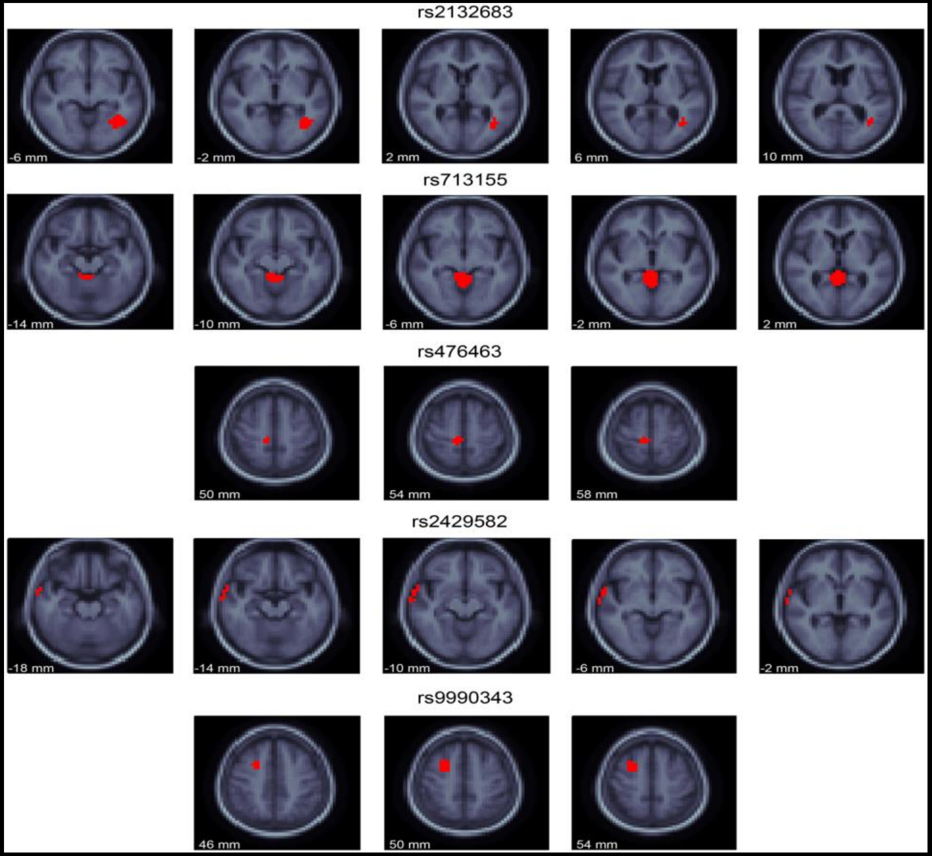
rs	chr	position	A1	A2	Region	WTCCC 1924 cases 2938 controls OR (95% CI)	P_{add}	Replication meta-analysis 3757 cases 5346 controls OR (95% CI)	P_{add}	All UK sample meta-analysis 5681 cases 8284 controls OR (95% CI)	P_{add}	DGI 6529 cases 7252 controls OR (95% CI)	P_{add}	FUSION 2376 cases 2432 controls OR (95% CI)	P_{add}	All combined 14,586 cases 17,968 controls OR (95% CI)	P_{add}
rs8050136	16	52373776	A	C	*FTO*	1.27 (1.16–1.37)	2.0×10^{-8}	1.22 (1.12–1.32)	5.4×10^{-7}	1.23 (1.18–1.32)	7.3×10^{-14}	1.03 (0.91–1.17)	0.25	1.11 (1.02–1.20)	0.017	1.17 (1.12–1.22)	1.3×10^{-12}
rs10946398	6	20769013	A	C	*CDKAL1*	1.20 (1.10–1.31)	2.5×10^{-5}	1.14 (1.07–1.22)	8.3×10^{-5}	1.16 (1.10–1.22)	1.3×10^{-8}	1.08 (1.03–1.14)	2.4×10^{-3}	1.12 (1.03–1.22)	9.5×10^{-3}	1.12 (1.08–1.16)	4.1×10^{-11}
rs5015480	10	94455539	C	T	*HHEX*	1.22 (1.12–1.33)	5.4×10^{-6}	–	–	1.13 (1.07–1.19)	4.6×10^{-6}	1.14 (1.06–1.22)	1.7×10^{-4}	1.10 (1.01–1.19)	0.025	1.13 (1.08–1.17)	5.7×10^{-10}
rs1111875	10	94452862	C	T	*HHEX*	–	–	1.08 (1.01–1.15)	0.020								
rs10811661	9	22124094	C	T	*CDKN2B*	1.22 (1.09–1.37)	7.6×10^{-4}	1.18 (1.08–1.28)	1.7×10^{-4}	1.19 (1.11–1.28)	4.9×10^{-7}	1.20 (1.12–1.28)	5.4×10^{-8}	1.20 (1.07–1.36)	2.2×10^{-3}	1.20 (1.14–1.25)	7.8×10^{-15}
rs564398	9	22019547	C	T	*CDKN2B*	1.16 (1.07–1.27)	3.2×10^{-4}	1.12 (1.05–1.19)	8.6×10^{-4}	1.13 (1.08–1.19)	1.3×10^{-6}	1.05 (0.94–1.17)	0.5	1.13 (1.01–1.27)	0.039	1.12 (1.07–1.17)	1.2×10^{-7}
rs4402960	3	186994389	G	T	*IGF2BP2*	1.15 (1.05–1.25)	1.7×10^{-3}	1.09 (1.01–1.16)	0.018	1.11 (1.05–1.16)	1.6×10^{-4}	1.17 (1.11–1.23)	1.7×10^{-9}	1.18 (1.08–1.28)	2.4×10^{-4}	1.14 (1.11–1.18)	8.6×10^{-16}
rs13266634	8	118253964	C	T	*SLC30A8*	1.12 (1.02–1.23)	0.020	1.12 (1.04–1.19)	1.2×10^{-3}	1.12 (1.05–1.18)	7.0×10^{-5}	1.07 (1.00–1.16)	0.047	1.18 (1.09–1.29)	7.0×10^{-5}	1.12 (1.07–1.16)	5.3×10^{-8}
rs7901695	10	114744078	C	T	*TCF7L2*	1.37 (1.25–1.49)	6.7×10^{-13}	–	–	–	–	1.38 (1.31–1.46)	2.3×10^{-31}	1.34 (1.21–1.49)	1.4×10^{-6}	1.37 (1.31–1.43)	1.0×10^{-48}
rs5215	11	17365206	C	T	*KCNJ11*	1.15 (1.05–1.25)	1.3×10^{-3}	–	–	–	–	1.15 (1.09–1.21)	1.0×10^{-7}	1.11 (1.02–1.20)	0.014	1.14 (1.10–1.19)	5.0×10^{-11}
rs1801282	3	12368125	C	G	*PPARG*	1.23 (1.09–1.41)	1.3×10^{-3}	–	–	–	–	1.09 (1.01–1.16)	0.019	1.20 (1.07–1.33)	1.4×10^{-3}	1.14 (1.08–1.20)	1.7×10^{-6}

GWAS for type II Diabetics:11 selected by the table out of ~400,000

Selection by a Figure

Goal: Association between volume changes at voxels with genotype Stein et al.'10)

1 ← Voxels searched → 32,000

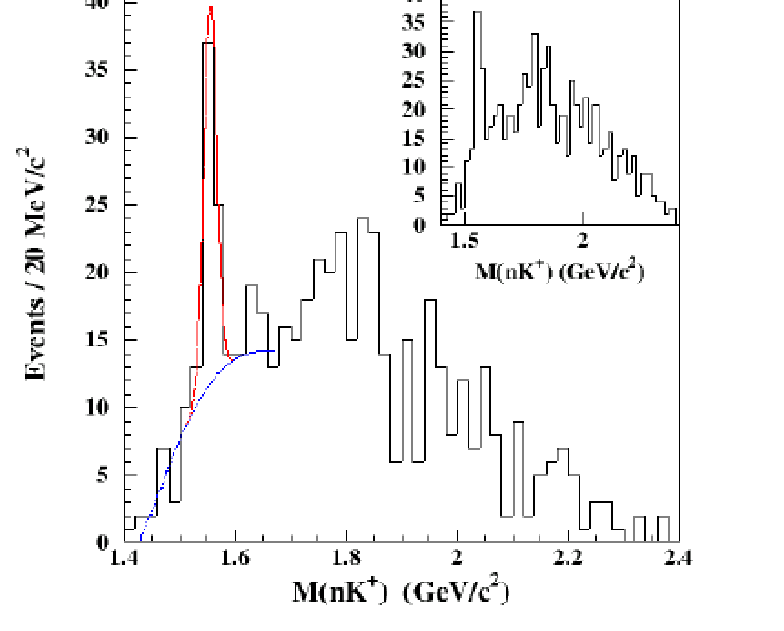


1 ↑ SNPs ↓ 448,000

number of tests ~ 13,000,000,000

Look Elsewhere Effect (LEE)

Sent to me by Louis Lyons



Prob of bgd fluctuation **at that place** = local p-value
Prob of bgd fluctuation 'anywhere' = global p-value
Global p > Local p
Where is `anywhere'?

a) Any location in this histogram in sensible range
b) Any location in this histogram
c) Also in histogram produced with different cuts, binning, etc.
d) Also in other plausible histograms for this analysis
e) Also in other searches in this PHYSICS group (e.g. SUSY at CMS)
f) In any search in this experiment (e.g. CMS)
g) In all CERN expts (e.g. LHC expts + NA62 + OPERA + ASACUSA + ….)
h) In all HEP expts
etc.

d) relevant for graduate student doing analysis
f) relevant for experiment's Spokesperson

INFORMAL CONSENSUS: Quote local p, and global p according to a) above. Explain which global p
N.B. Needs lots of MC to determine (global) p-value
Assyptotics enable extrapolation from lower significance {Gross and Vitells EPJ **C70**(2010) 525}

Why 5σ for Discovery?

Statisticians ridicule our belief in extreme tails (esp. for systematics)

Our reasons:

1) Past history (Many 3σ and 4σ effects have gone away)

2) LEE

3) Worries about underestimated systematics

4) Subconscious Bayes calculation

$$\frac{p(H_1|x)}{p(H_0|x)} = \frac{p(x|H_1)}{p(x|H_0)} * \frac{\pi(H_1)}{\pi(H_0)}$$

Posterior Likelihood Priors

prob ratio

“Extraordinary claims require extraordinary evidence”

N.B. Points 2), 3) and 4) are experiment-dependent

Alternative suggestion:

L.L. “Discovering the significance of 5σ” http://arxiv.org/abs/1310.1284

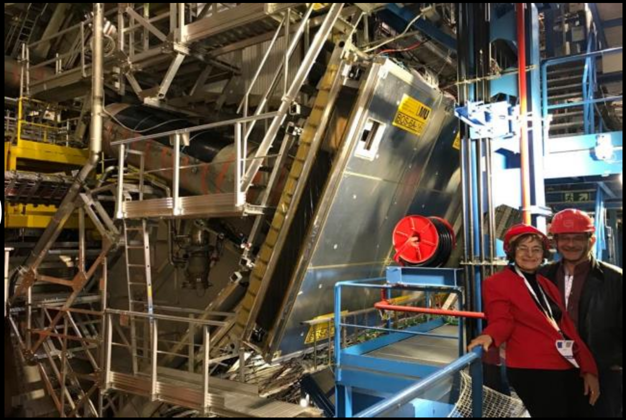
The industrialization of the scientific process

1888

1999

1950

2010



4. Addressing selective inference

A. Simultaneous over all possible selections (SoP)

B. Simultaneous over the selected (SoS)

C. Conditional over the selected (CoS)

D. On the average over the selected (AOS)

A. **Simultaneous** over all possible selections

The *FamilyWise error-rate (FWER) :*

For testing H_i's: R_i *=1* if H_i rejected V_i*=1* if rejected in error; otherwise *0*

$R=\Sigma R_i$ is number rejected $V=\Sigma V_i$ rejected in error

$$Pr(V \geq 1) \leq \alpha$$

For CIs : $Pr(\exists\ i,\ \mu_i \notin CI_i(Y)) \leq \alpha$ $=Pr(V \geq 1) \leq \alpha$

For any $S(\mathbf{Y}) \subset \{1,2,\ldots m)$ the same properties hold

Often very conservative

Pairwise Comparisons - 1950's Tukey, Scheffe

The Bonferroni simultaneous procedure

If we test each hypothesis separately at level α_{BON}

$$E(V)=E(\Sigma V_i) = \Sigma E(V_i) \le m_0\, \alpha_{BON} \le m\ \alpha_{BON}$$

To assure $E(V)\le\alpha$

we may use. $\alpha_{BON} = \alpha/m$ as a threshold

Implying $FWER = Prob(V \ge 1) \le E(V) \le \alpha$

$\widehat{m_0}$

Bonferroni adjusted p-value is $p_{BON,i}=mp_i$

Random Field based thresholding

Adler (1981), Worsley & Friston (fMRI, ‘96)

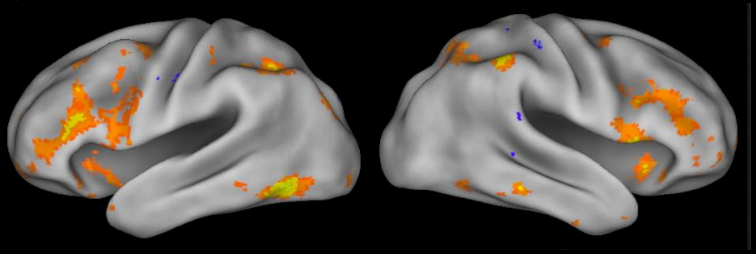
If the signal is in a smooth Gaussian field

- *Pr(max z(s)>t)* is a function of the Euler Characteristic

Or if signal is also smooth

- Using the special extent of level cluster S, say H=hight*width

$Pr(H(S)>h\,|\,z(s)>t) \sim exp\{-t(s/c)^{2/d}\}$



Then use Bonferroni over all such S.

Topological Data Analysis

Use in fMRI raised problems of replicability

The False Discovery Rate (FDR) criterion

Benjamini and Hochberg (95)

R = # rejected hypotheses = # discoveries

V of these may be in error = # false discoveries

The error (type I) in the entire study is measured by

$$FDP = V/R \qquad R > 0$$

$$= 0 \qquad R = 0$$

i.e. the proportion of false discoveries among the discoveries

$$FDR = E(FDP)$$

The goal: Maximize R while controlling $FDR \le q$

Does it make sense?

- Inspecting 100 features:

2 false ones among 50 discovered - *bearable*

2 false ones among 4 discovered - ***unbearable***

So this error rate is adaptive

- The same argument holds when inspecting 10,000

So this error rate is scalable

- If nothing is "real" controlling the FDR at level *q* guarantees

$$Prob(\ V \geq 1\ ) = E(\ V/R\ ) = FDR \leq q$$

- But otherwise

$$Prob(\ V \geq 1\ ) \geq FDR$$

So there is room for improving detection power

The BH procedure

Benjamini-Hochberg 1995 (BH)

Let P_i be the observed p-value of the test for H_i

- Order the p-values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$
- Let

$$k = \max\{i : p_{(i)} \pounds (i/m)q\}$$

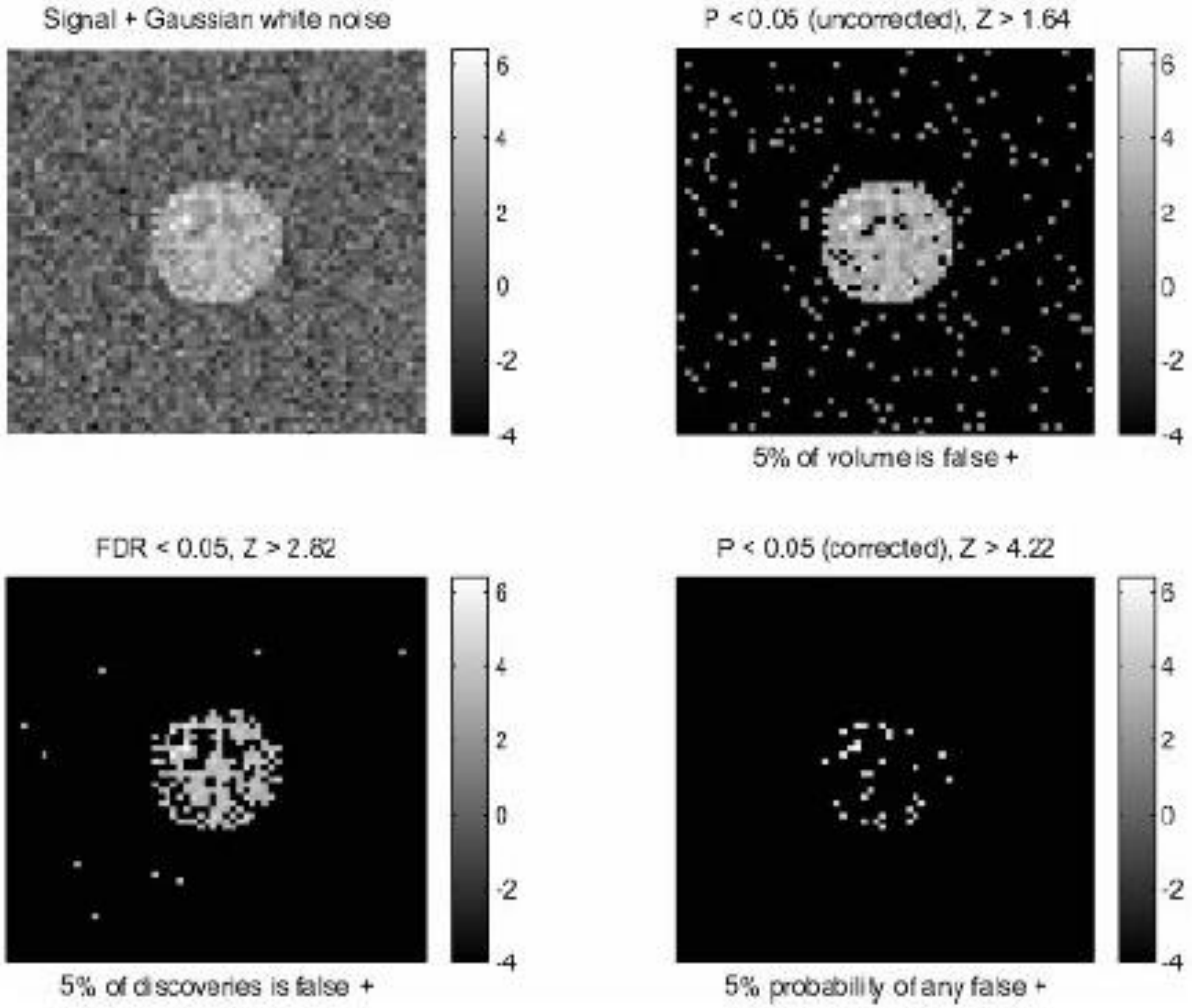
- Reject

$$H_{(1)}, H_{(2)}, \ldots, H_{(k)}$$

And in adjusted p-value form

$$p^{BH}_{(i)} = min \{ p_{(j)}m/j,\ \ j \geq I ; 1 \}$$

and reject if $p^{BH}_{(i)} \leq q$. These are now called q-values.



Signal + Gaussian white noise

P < 0.05 (uncorrected), Z > 1.64

5% of volume is false +

FDR < 0.05, Z > 2.82

5% of discoveries is false +

P < 0.05 (corrected), Z > 4.22

5% probability of any false +

Figure 3: Illustration of the difference between False Discovery Rate and Bonferroni/random field methods for thresholding an image.

- If only one strong signal $p_1 \leq q/m$

  as strict as Bonferroni

- If many signals are strong threshold for significance close to q

  i.e. very small penalty for addressing selection – large gain.

- The gain is still large if the signal is sparse m_0/m -> 1
- If the test statistics are

  independent

  positive regression dependent one sided

  FDR $\leq (m_0/m)q$

  General dependence

  FDR $\leq (m_0/m)q(1+1/2+1/3+\ldots+1/m)$

  So use BH with $q/log(m)$ (BY procedure)

Adaptive procedures that control FDR

Recall the $\boldsymbol{m_0/m}$ (=p_0) factor of conservativeness

Hence: if $\boldsymbol{m_0}$ is known, the BH procedure with

q*=$\boldsymbol{q}$ $\boldsymbol{(m/m_0)} \geq \boldsymbol{q}$ controls the FDR at level q exactly

i.e. an "FDR Oracle"

The essence of adaptive procedures:

Estimate $\boldsymbol{m_0}$ (or p_0) from the p-values - from the large ones

They can gain power when the signal is dense $\boldsymbol{(m_0/m) < 1}$

Schweder&Spjotvol ('86), Hochberg&BY ('90), BY&Hochberg ('00)

`Parametric modeling; EM algorithm with mixtures; Ratio of densities at 0, Spectral analysis; Histogram analysis,…

Resampling procedures

Yekutieli & BY ('99) Efron et al ('01), Storey('01), Storey & Tibshirani ('03) Genovese & Wasserman ('04) Troendle et al …van der Laan & Dudoit ('09)

Empirical Bayes (local false discovery)

Efron ('03) … Efron's book Large Scale Inference ('10)

Model selection with FDR penalties

Abramovich,YB, Donoho &Johnstone ('10)

Knockoff procedures for model selection with FDR control

Candes & Foygel-Barber ('15) Wald Lecture JSM '17

Model selection while addressing selection

Microarray dataset of 10 normal and 86 cancerous lung tissues (Beer, et al., ‘02), 7127 features,
analyzed in Rupin’s Lab (Bionformatics, ’05)

The goal: Produce a stable ranked gene list,

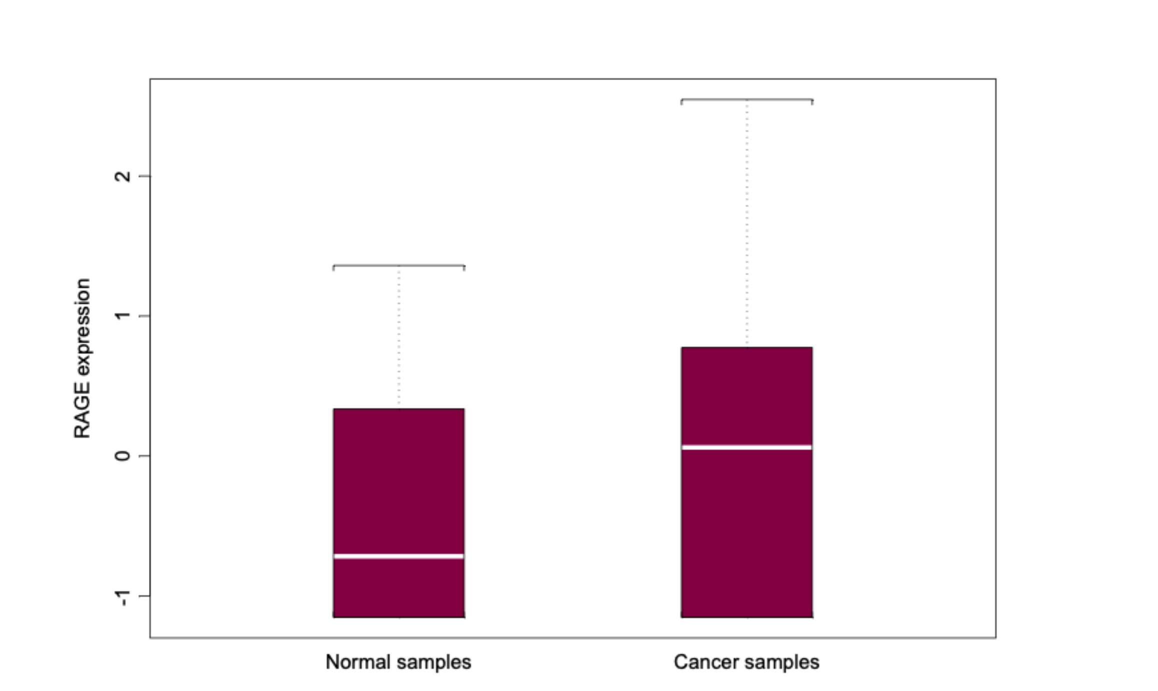
the top of which should be a “good” set of classifiers to build on.

Rupin's Lab Method:

(i) Producing 1000 different gene sets according to the SVM models of sizes 5 up to 100, on bootstrapped samples

(ii) ranking the genes according to their repeatability frequency in the ensemble of predictive gene sets.

Result: The gene with the highest score was "Rage", its boxplot by two classes is presented below



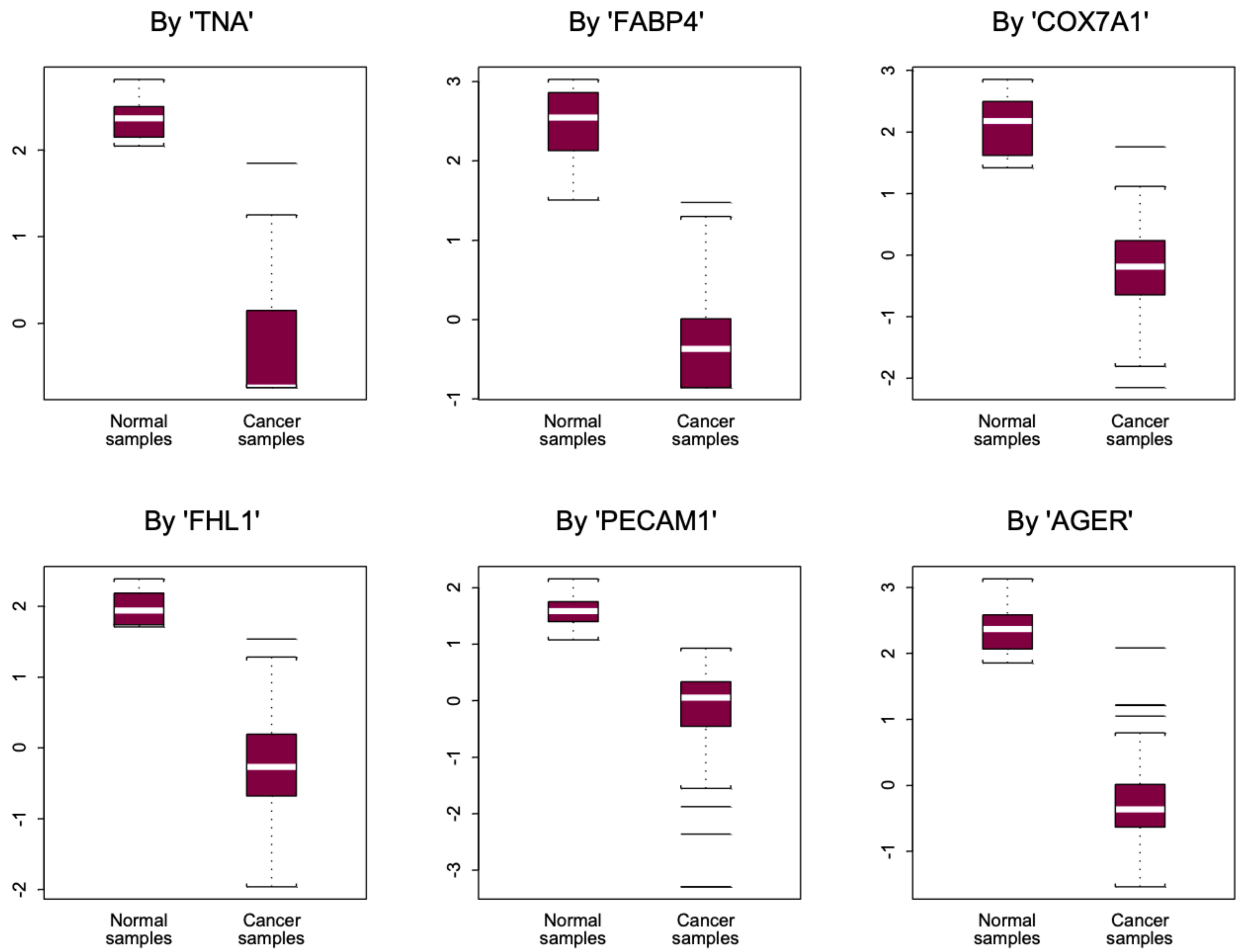
Selection adjusted regression

Forward (greedy) selection the features to enter the logistic model in order to minimizes the deviance plus FDR penalty.

Unlike the penalties in AIC, BIC or Cp that are linear in model size k ; but penalty per parameter unaffected by, the FDR penalty per parameter increases in size of the pool of potential features m and decreases in k. $\sim k*2\ln(\alpha m/k)$ YB & Gavrilov ('13)

- Replicating 120 times by bootstrapping,

In all replications only one gene is selected.


By 'TNA'
0
1
2
Normal samples
Cancer samples
By 'FABP4'
-1
0
1
2
3
Normal samples
Cancer samples
By 'COX7A1'
-2
-1
0
1
2
3
Normal samples
Cancer samples
By 'FHL1'
-2
-1
0
1
2
Normal samples
Cancer samples
By 'PECAM1'
-3
-2
-1
0
1
2
Normal samples
Cancer samples
By 'AGER'
-1
0
1
2
3
Normal samples
Cancer samples

A Flexible Approach

The BH procedure

Can be modified to reflect increasing detection power for some hypotheses at the expense of other ones
By introducing ν_i for each hypothesis, so $\Sigma\nu_i=m$

and $p^*_i=p_i/\nu_i$

The FDR criterion

Can be modified to reflect varying importance of error
By introducing w_i for each hypothesis, so $\Sigma w_i=m$

$$FDP_w=(\Sigma w_i V_l)/(\Sigma w_i R_l)$$

Thereby can be modified to reflect price of follow-up studies, areas of bumps, …

D. On the average over the selected

Rephrase the False Discovery Rate (FDR) for testing:
*S(**Y**)* selects the rejected hypotheses; *R= |S(**Y**)|*
V is the number in S(Y) of type I errors

So $FDP = V/R = (\sum_{i \varepsilon S(Y)} V_i) / |S(Y)|$ *if R>0*

$= 0$ *if R=0*

And

$$FDR = E(FDP)$$

FDR is the expected average # errors over the selected

For the False Coverage-statement Rate (FCR) :

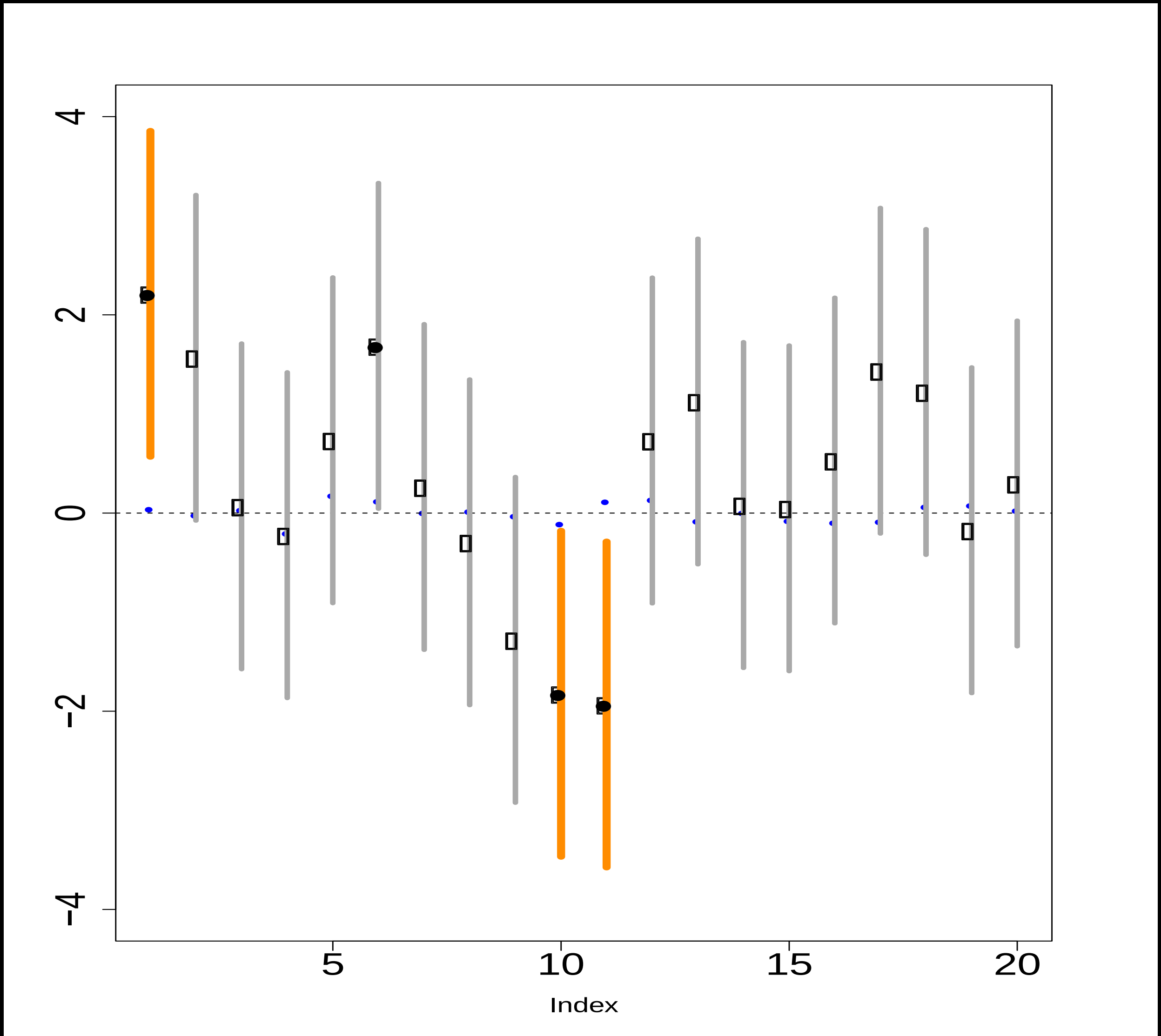
Set V_i=1 for a selected non-covering interval

$$FCR=E(\sum_{i \varepsilon S(Y)} V_i) / |S(Y)|)$$

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.

General FCR controlling CIs

Selecting from *m* features the ‘interesting ones’

If selection is ‘simple’

For *each selected one*

construct a marginal $1-q\frac{|S(Y)|}{m}$ Conf. Intervals

YB and Yekutieli ’05

Beyond Positive Regression Dependence?

Recognizing a family

A family is

The smallest set of items of inference in an analysis,

From which any selection of results for presentation and highlighting could be made,

And be as useful.

Exchangibility in meaning

Different researchers can have different goals and thus define differently the families – still decisions can be defendable and with no arbitrariness.

Recognizing a family

In a report of a clinical trial, not all hypotheses tested are a single family. There are at least 3 families:

- Comparisons of baseline characteristics
- Comparisons of endpoints capturing treatment effects
- Comparisons of safety endpoint

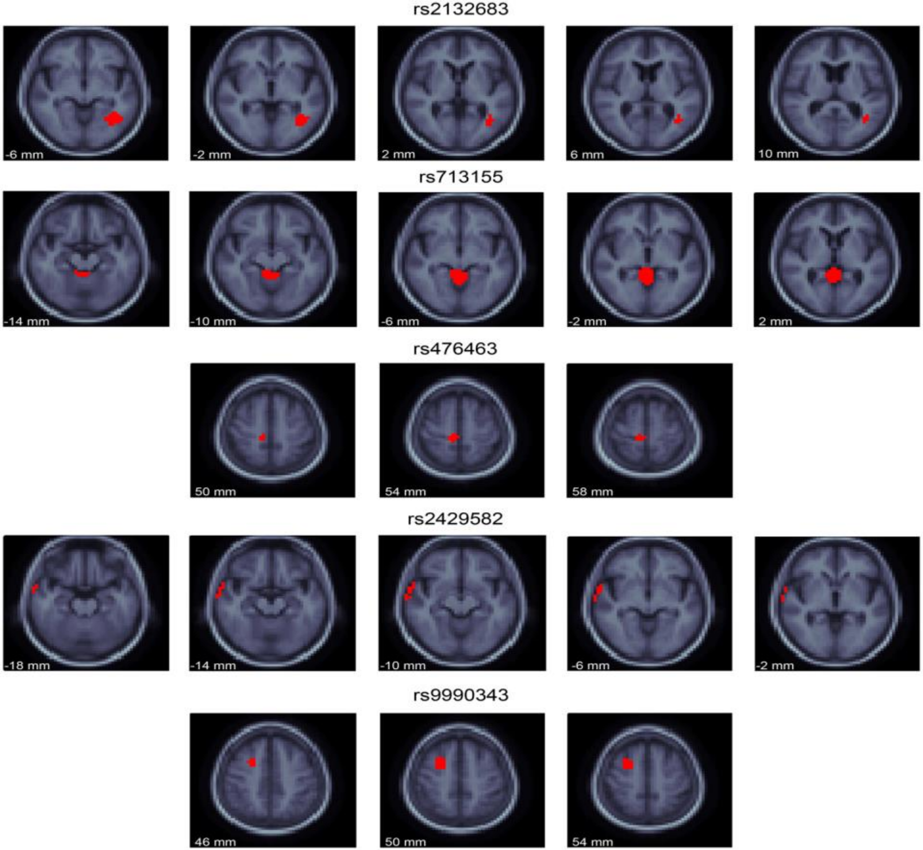
A selected finding from one family cannot play the role of a finding from another family

Note: A family is not defined by the statistical dependency structure of the inferences.

D.2 Hierarchical testing of family of families

Goal: Association between volume changes at voxels with genotype Stein et al.’10)

1 ← Voxels searched 32,000



1 ↑ SNPs ↓ 448,000

number of tests ~ 13,000,000,000

Selection adjusted testing of families

Let H_{ij} be the the hypotheses in family F_i, $j=1,..m_i$; $i=1,\ldots,m$

with $\boldsymbol{Y}=\{Y_{ij}\}$ or with p-values $\boldsymbol{P}=\{p_{ij}\}$)

$S(\boldsymbol{Y})$ is a selection procedure of families.

$|S(\boldsymbol{Y})|$ the (random) number of families selected.

The control of error FDP

on the average over the selected families means

$$E\left(\frac{\sum_{i\in S(P)} FDP_i}{|S(P)|}\right) \leq q$$

BH over all hypotheses may be too liberal on the family level!

(BH-q, BH-q**R**/m) - hierarchical testing

Test the intersection hypothesis $\cap_j H_{ij}$ in F_i , using Simes test

$$p_i^* = min(\ p_{i(j)} m_i / j)$$

Which is also minimum of the BH adjusted p-value in a family

$$q_{i(j)}^{BH} = min\ _{k>j}\ (p_{i(k)} m_i / k)$$

and

$$p_i^* = min(\ q_{i(j)}^{BH})$$

(BH-q, BH-q**R**/m) - hierarchical testing

Test the families using BH-q with p_i* ; select the rejected ***R***.

Within each selected family use BH at level *q(**R** / m)*

(1) $$E\left(\frac{\sum_{i\in S(P)} FDP_i}{|S(P)|}\right) \le q$$

(2) $FDR \le q$ within families;

Conditions are as needed for the BH

YB&Bogomolov ‘14

Results for Association between volume changes at voxels with genotype

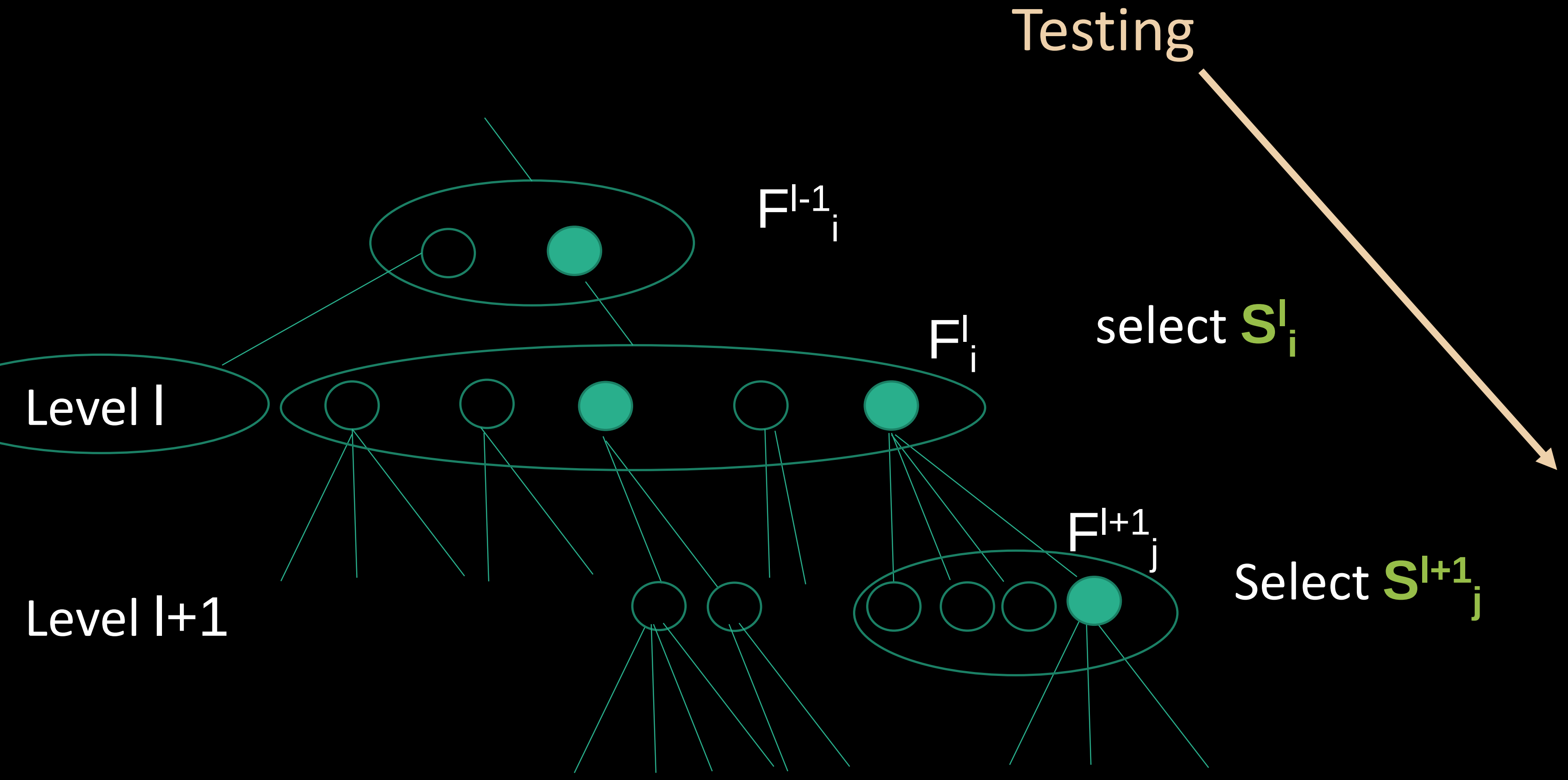
- Family = the set of all association hypotheses for a specific SNP and all voxels (~34K)

  Calculate p-value per SNP-family testing "is there something at all".

- Select SNPs while controlling FDR over SNPs: 35 SNPs
- Test voxels within families of selected SNPs, assuring FDR control on the average over the selected – using BH at level .05*35/448,000
- For most SNPs ≤ 50 voxels; the max 400 voxels.

L levels in the tree

The general hierarchical structure



L levels in the tree

The recursive error-rate at each level

H_j at level k

$$FDP(\mathrm{F}^{k}_{j}) = \frac{\sum_{r\in \mathbf{S}^{k+1}_{j}} FDP\,(\mathrm{F}^{k+1}_{r})}{|\mathbf{S}^{k+1}_{j}|}$$

Errors

$$\mathrm{sFDP}^{L} = FDP(\mathrm{F}^{0})$$

Stopping the testing at any level l, we start the recursion from l

$$\mathrm{sFDR}^{l} = \mathrm{E}(\ \mathrm{sFDP}^{l}\ )$$

The expected hierarchically averaged *FDP* in higher levels

L levels in the tree

The sFDR^l error-rate at level l : The expected hierarchically averaged *FDP* in higher levels

Another interpretation for sFDR^l = E(sFDP^l)

$$\text{sFDP}^l = \sum_{j\ s.t.\ \text{F}^l_j\ is\ tested} w_j^l FDP(\text{F}^l_j)$$

$$w_j^l = \Big[\prod^{Ancestors\ of(\text{F}^l_j)} |\text{S}^k_i|\Big]^{-1}$$

The more extreme is the selection that leads to a hypothesis the larger the weight its error gets

Gene-expression association with its nearby SNPs in multiple tissues

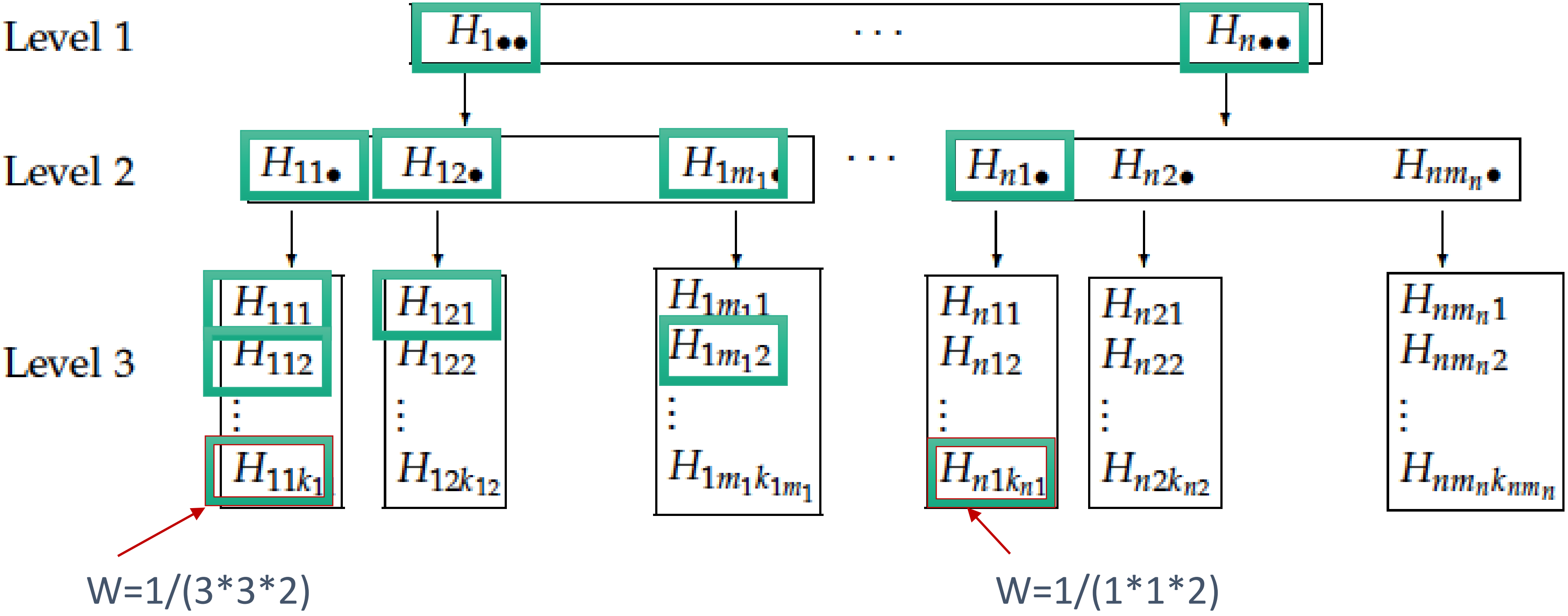


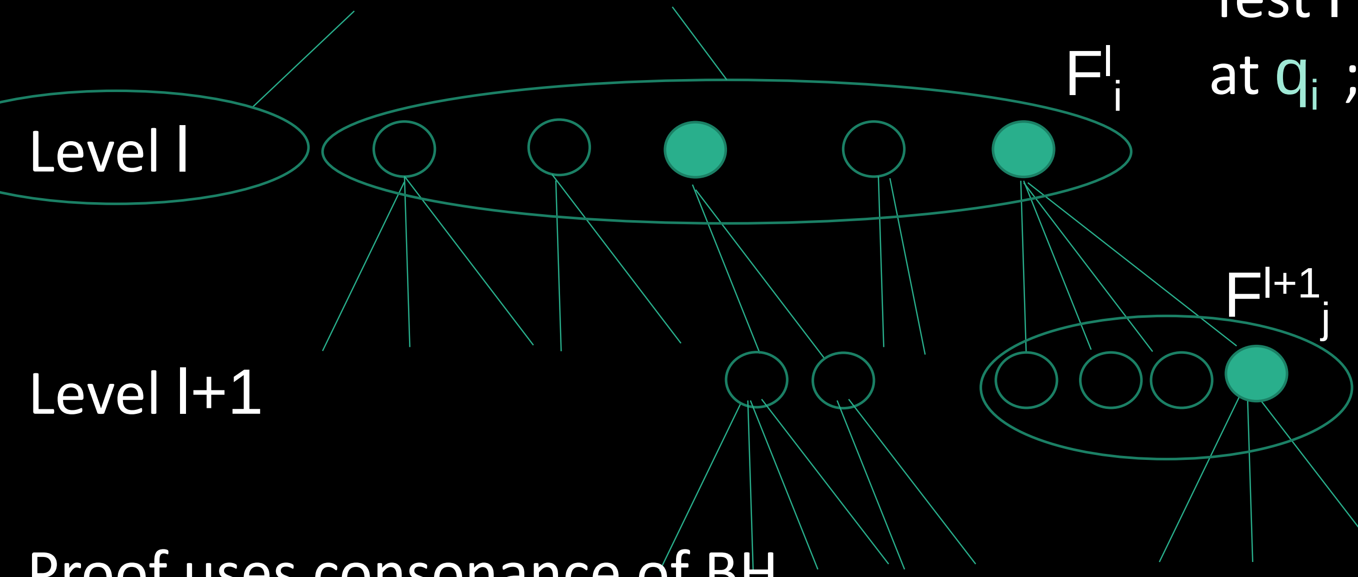
Figure 1: *Hierarchical structure of the hypotheses*

L levels in the tree: The TreeBH procedure

Using it, for any $l \leq L$ $\mathrm{sFDR}^l \leq q$.

Testing

The TreeBH procedure:



Level l

Level l+1

F^l_i

F^{l+1}_j

Test F^l_i with BH at q_i ; select $\mathbf{S}^l_i$

Test F^{l+1}_j with BH

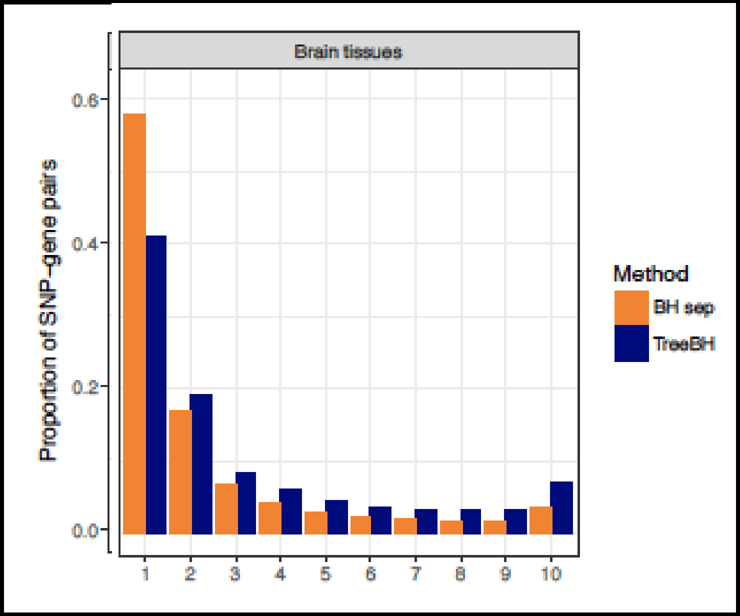
at $q_j = \frac{|\mathbf{S}^l_i|}{|F^l_i|} q_i$

Select $\mathbf{S}^{l+1}_j$

Proof uses consonance of BH

eQTL in multiple tissues - TreeBH

Proportion of gene-SNP pairs



BH- separately per Tissue

TreeBH

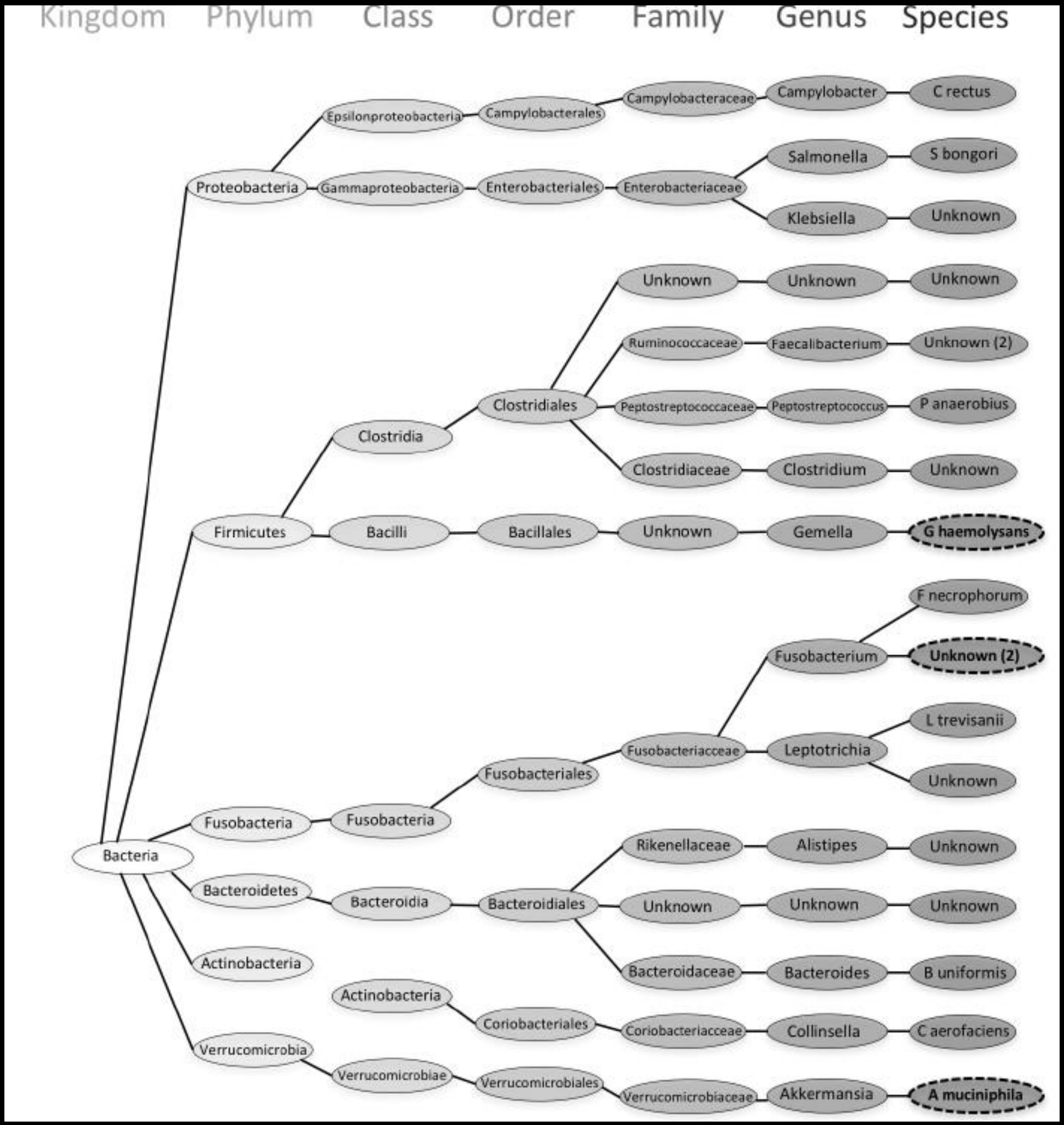
Number of tissues sharing gene-SNP pairs

Association of gut microbiome with Colon Cancer

- 496 microorganisms N= 177 (86 tumors)
- Abundance determined by rDNA & compared between cancer and normal

Offers on the average over the selected inference at the levels of Phylum, Class, Order, …. Species

Bogomolov, Peterson, YB Sabbati ('17+)



Error-rates for selective inference

A. *Simultaneous over all possible selections (SoP)*

B. *Simultaneous over the selected (SoS)*

C. *Conditional over the selected (CoS)*

D. *On the average over the selected (AOS)*

C. Conditional over the selected

Selecting from a set of features by a selection rule *S(**Y**)*

For *each one*

construct a marginal conditional confidence interval

$$Pr(\mu_i \notin CI_i(\mathbf{Y}) \mid i \in S(\mathbf{Y})) \leq \alpha$$

E.g. Select the largest one; Bigger than 2; p-value ≤.01 ;

Coefficients in the Lasso

Conditional inference => **FDR**/FCR

Utilizing the selection procedure used

Select μ_i if its estimator is big enough

$$X_i = (Y_i \mid |Y_i| \geq c),$$

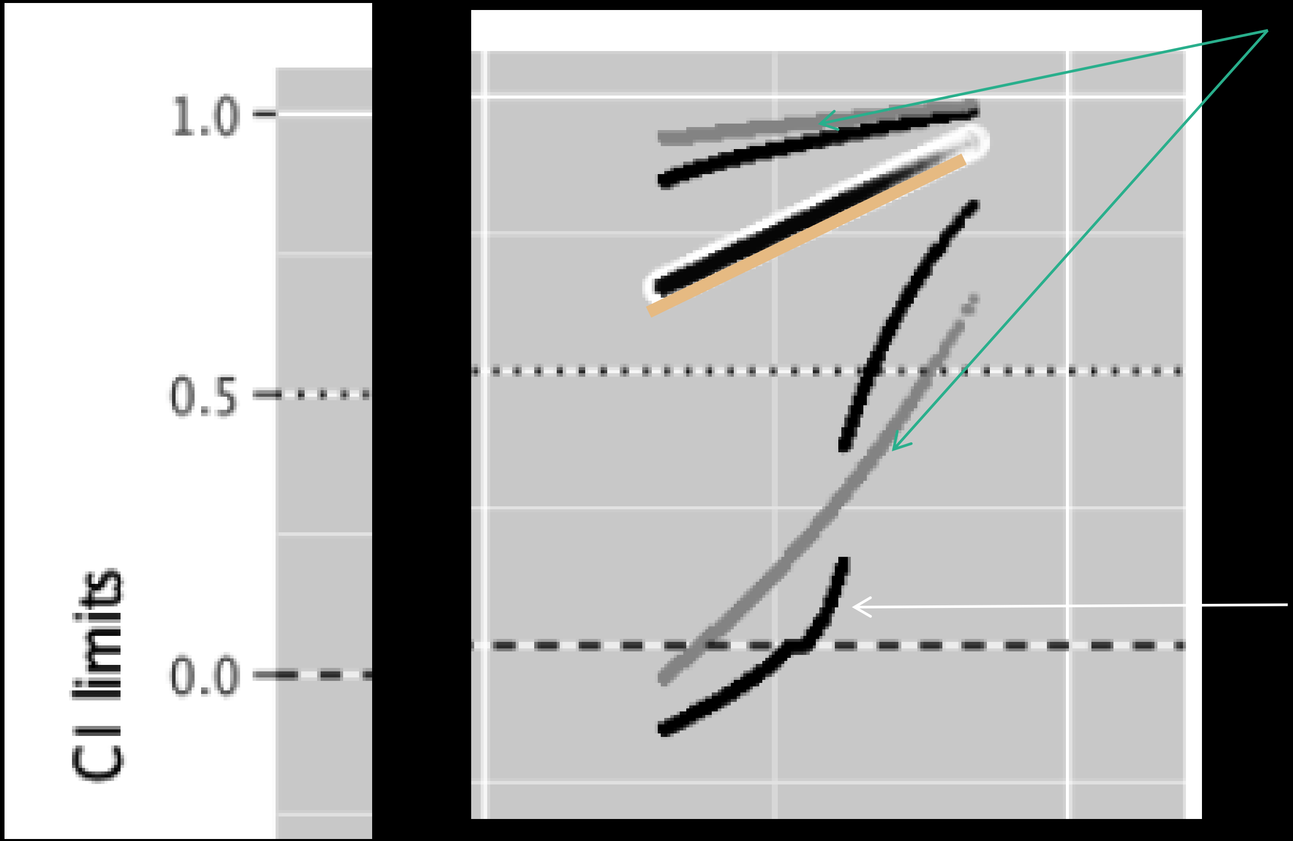
where c is fixed, say $z_{1-\alpha/2}$

or (simple) data dependent $c(\boldsymbol{Y})$.

Conditional density -> Acceptance region for each parameter (non-equivariant) with short 0-crossing -> inverting to get Conditional CIs -> offers FCR

Hedges ('84) for meta-analysis, Zhong &Prentice ('08) asymptotic dist'n in GWAS, Weinstein Fithian YB ('13)

FCR CIs vs Conditional ones

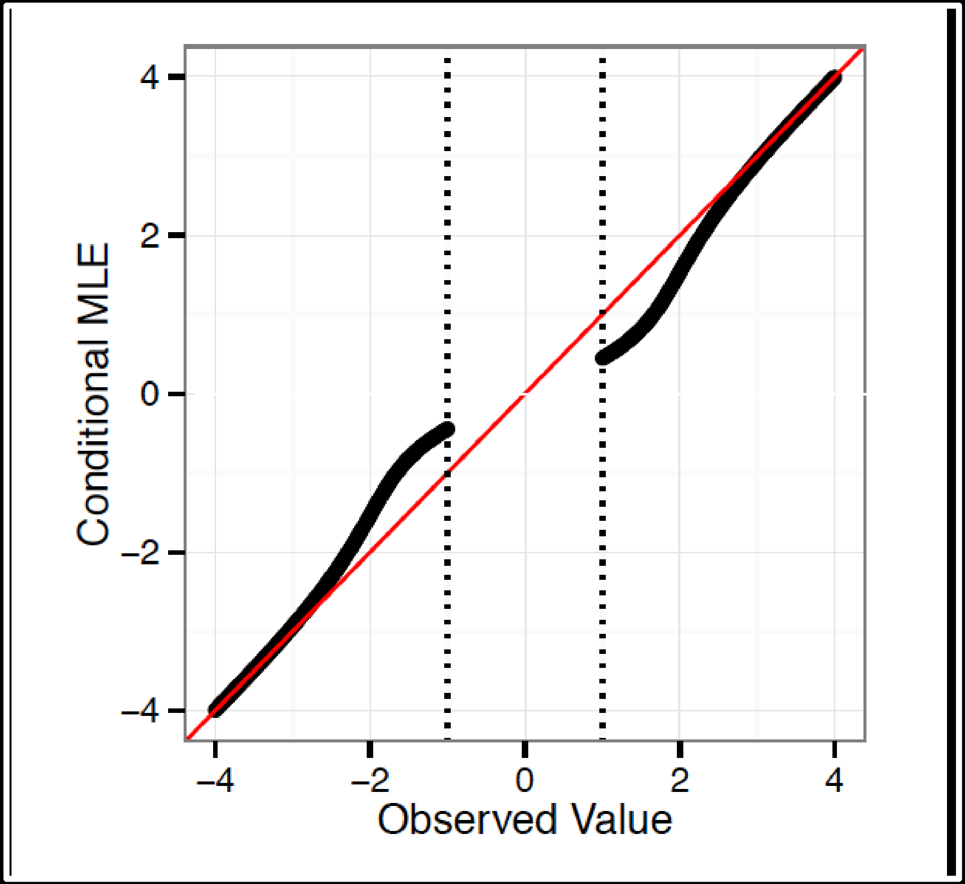


FCR

Conditional

Rosenblatt &YB (Neuroimage ‘14)

Conditional MLE



Hedges '84, Zhong and Prentice '08, Fithian, Sun, Taylor (16) YB and Meir (16+)

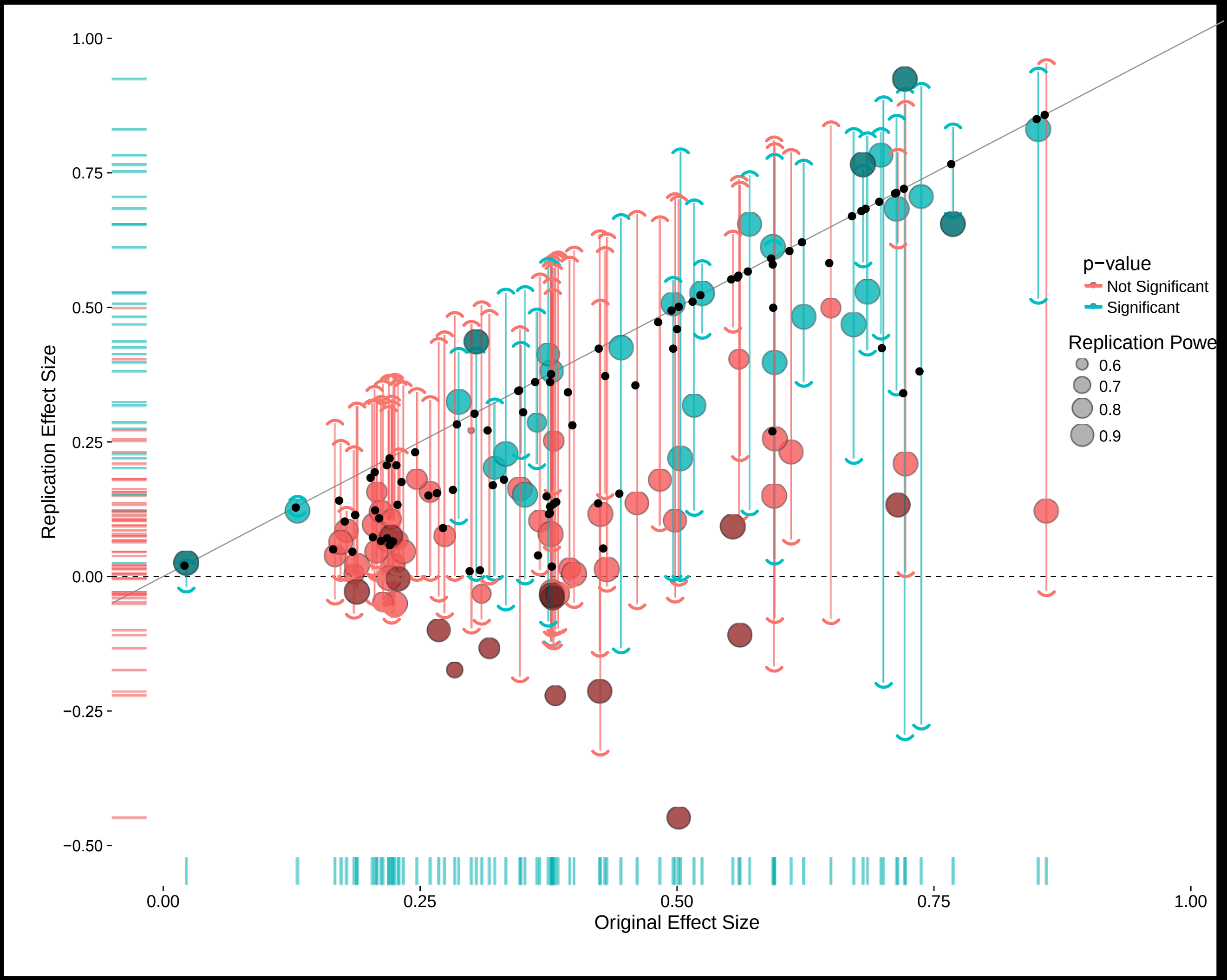
Both can be used to address 'publication bias'

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

p-value ≤ 0.05



77% fall in Cond. CI Instead of 47%

Inference on bumps

Benjamini Yuval, Taylor & Irrizari ('18)

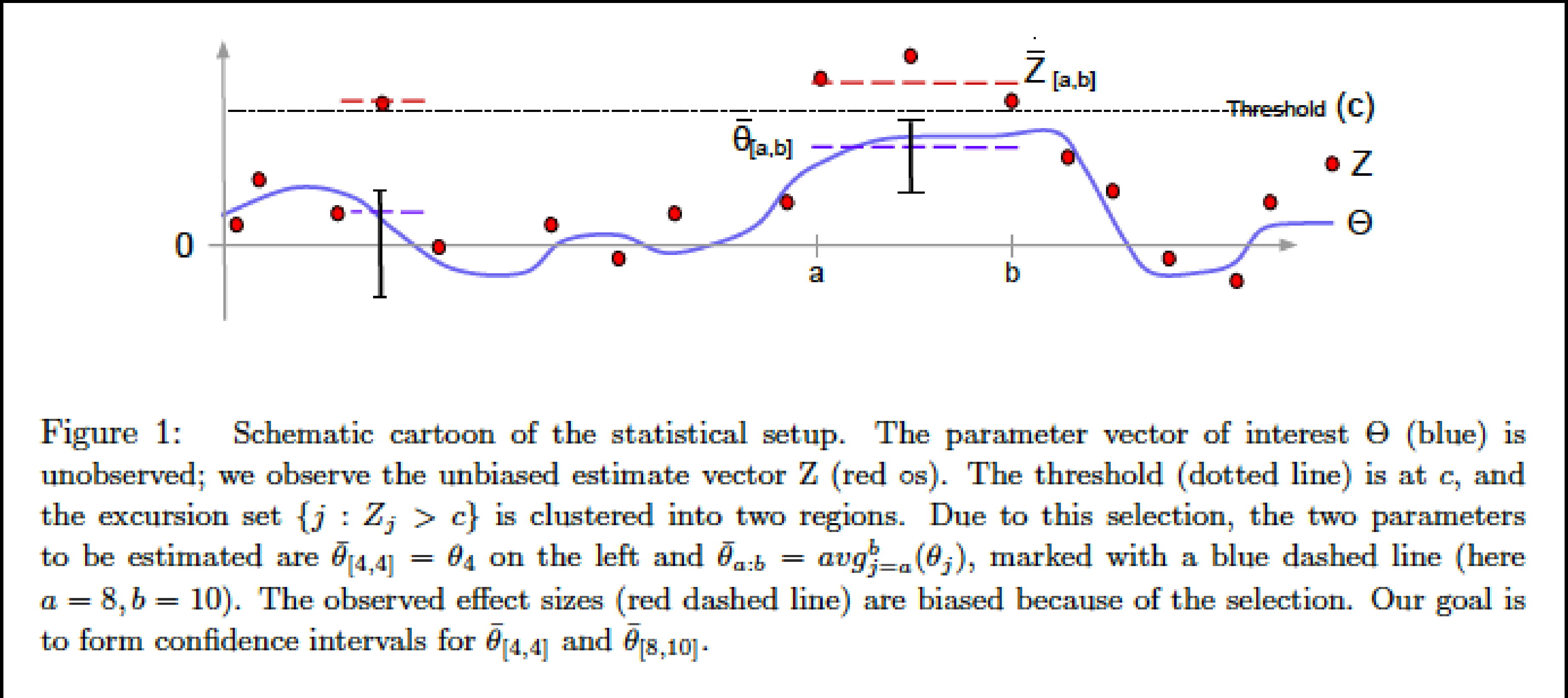
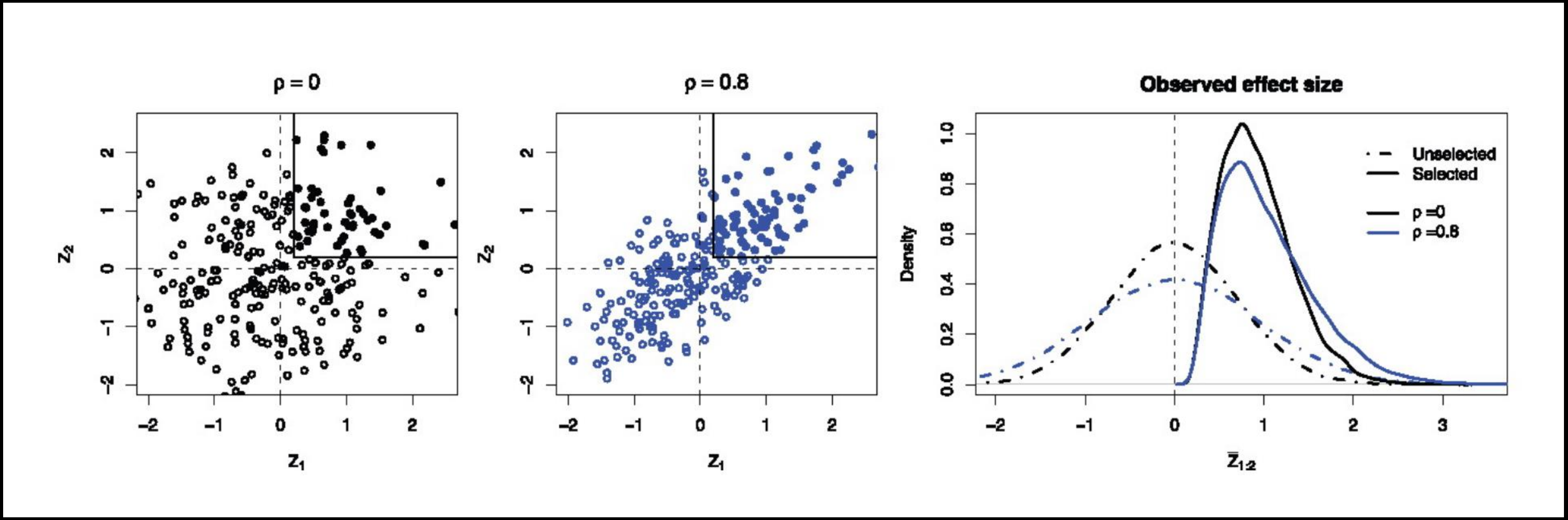


Figure 1: Schematic cartoon of the statistical setup. The parameter vector of interest Θ (blue) is unobserved; we observe the unbiased estimate vector Z (red os). The threshold (dotted line) is at c, and the excursion set $\{j : Z_j > c\}$ is clustered into two regions. Due to this selection, the two parameters to be estimated are $\bar{\theta}_{[4,4]} = \theta_4$ on the left and $\bar{\theta}_{a:b} = avg_{j=a}^{b}(\theta_j)$, marked with a blue dashed line (here $a = 8, b = 10$). The observed effect sizes (red dashed line) are biased because of the selection. Our goal is to form confidence intervals for $\bar{\theta}_{[4,4]}$ and $\bar{\theta}_{[8,10]}$.

- What’s the problem



Recent and ongoing work

On conditional inference for parameters after selecting a model with Lasso, forward selection, …

J. Taylor with coauthors and students (’13+):

Lockhart, Taylor, Tibshirani, R Tibshirani R, Lee, Dennis Sun, Yuekai

;Fithian and Wang (17+)

On hierarchical methods

Foygel-Barber, Ramadas , Chen, Wainright , Jordan (‘20)

On combination of the two

Heller, Meir, Chattergee, Krieger (18+)

Addressing inference after selection
In a database

If you torture your data long enough it will confess

similarly

If enough researchers (postdocs) torture your database

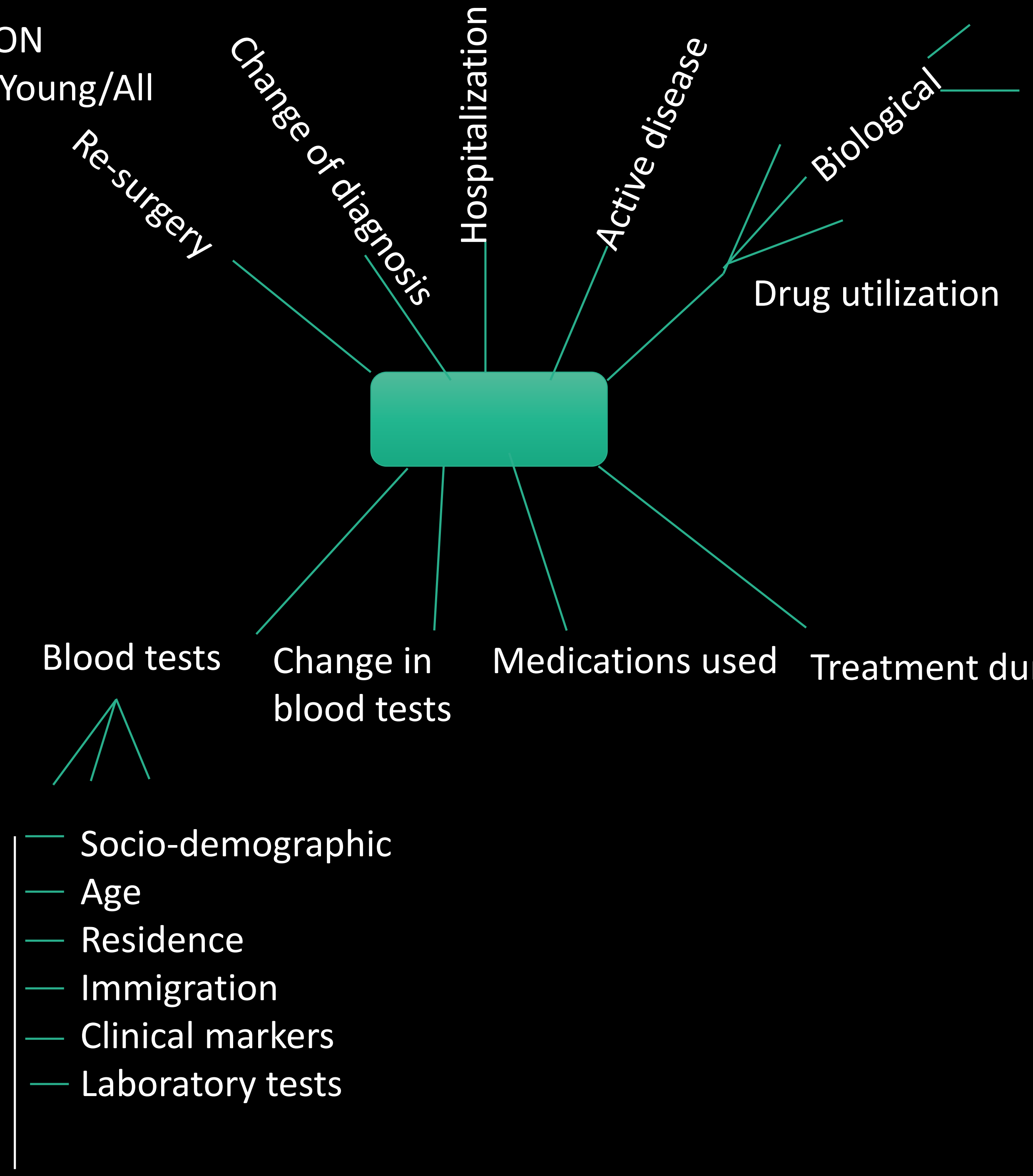
it will confess

A concrete example

Data collected from Israeli HMO about all patients in Israel with gut diseases.

Intended to serve as a database for studies by others

From some proposed protocols for studies investigating post-surgery, we could figure a structure



SUBPOPULATION

Old/Young/All

OUTCOMES

Re-surgery

Change of diagnosis

Hospitalization

Active disease

Biological

Drug utilization

PREDICTORS

Blood tests

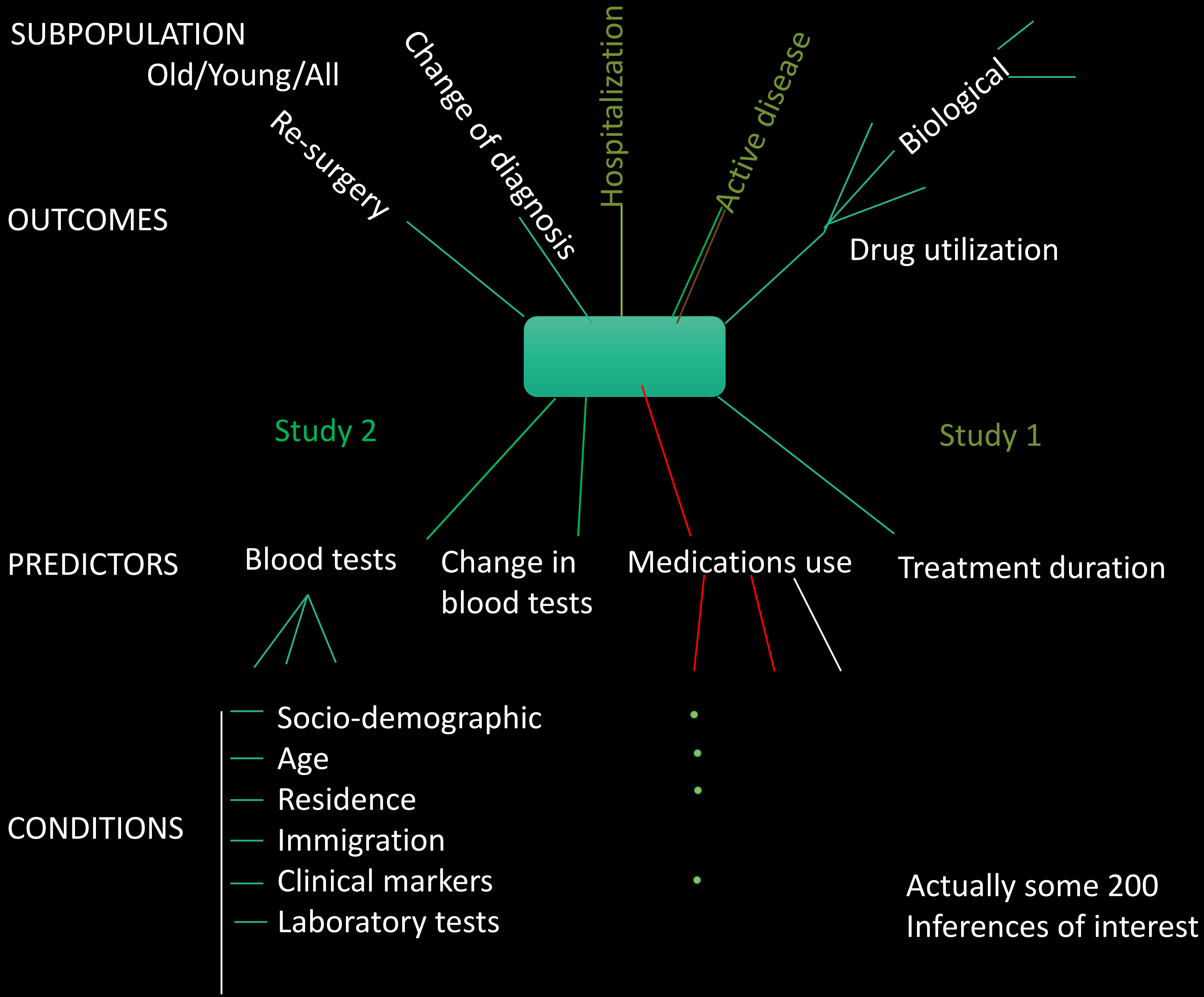
Change in blood tests

Medications used

Treatment duration

CONDITIONS

- Socio-demographic
- Age
- Residence
- Immigration
- Clinical markers
- Laboratory tests



SUBPOPULATION

Old/Young/All

OUTCOMES

Re-surgery

Change of diagnosis

Hospitalization

Active disease

Biological

Drug utilization

Study 2

Study 1

PREDICTORS

Blood tests

Change in blood tests

Medications use

Treatment duration

CONDITIONS

- Socio-demographic
- Age
- Residence
- Immigration
- Clinical markers
- Laboratory tests

Actually some 200
Inferences of interest

Emerging approach

- Require reporting in protocol as wide as possible array of questions of interest. Yet allow unplanned follow-ups.
- Allow designating questions of prime importance
- Require each researcher to adjust for selective inference.
  - Allow different weighing for prime ones
  - Allow hierarchical approach (including unplanned
- Tag consistently outcomes, interventions, conditions & populations across studies
- Deposit tagged results at the database

Emerging approach

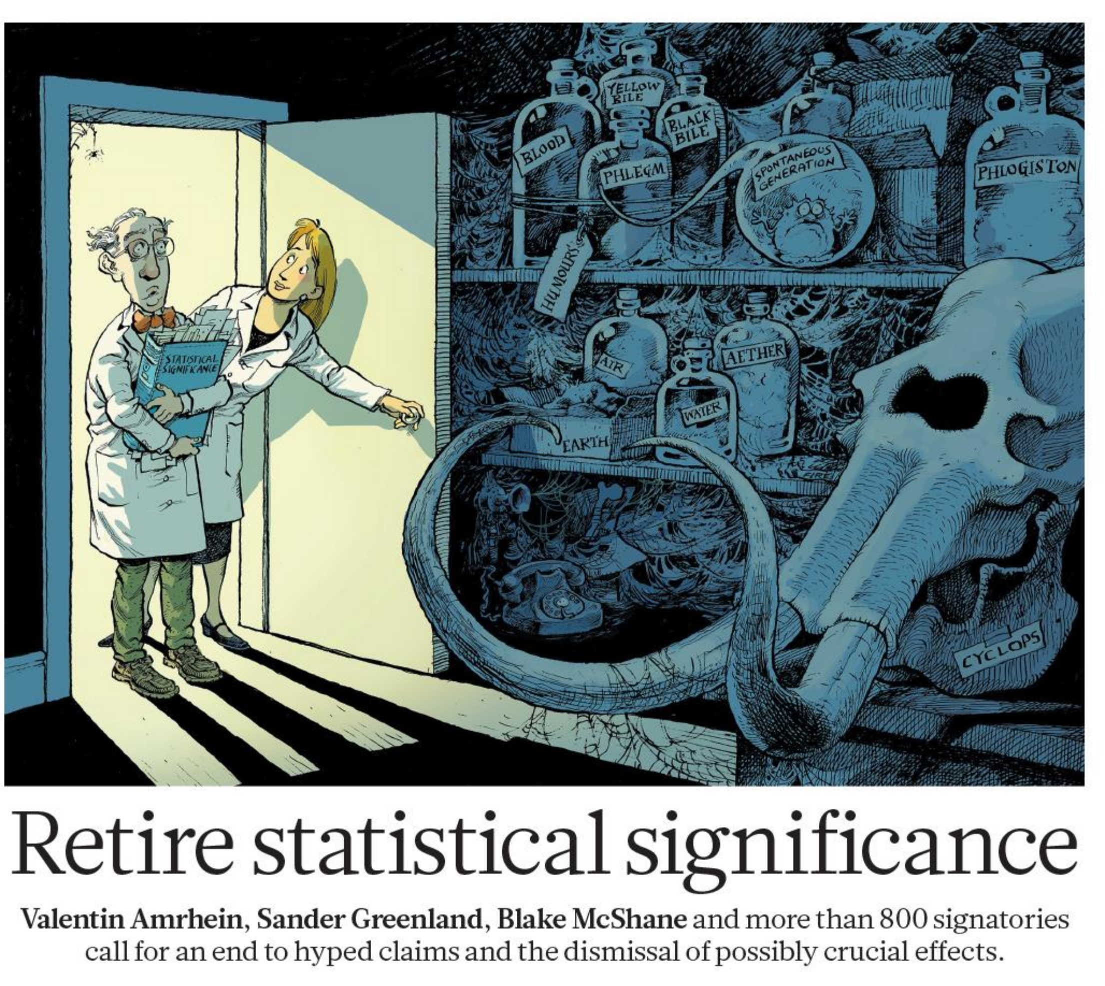
- Use meta-analysis retrospectively
- Only exchangeable-in-meaning questions should be adjusted for selection – but even across studies
- Can be adjusted hierarchically
- Check for stronger replicability and generalizability of results across sub-populations

It should it be the database management responsibility
To carry out such retrospective studies in order to assure the scientific integrity

The Status: Nature Magaszine

'Scientists rise up against statistical significance'

But also
‘~~confidence intervals~~’
-> ‘plausibility intervals’



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 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?"*

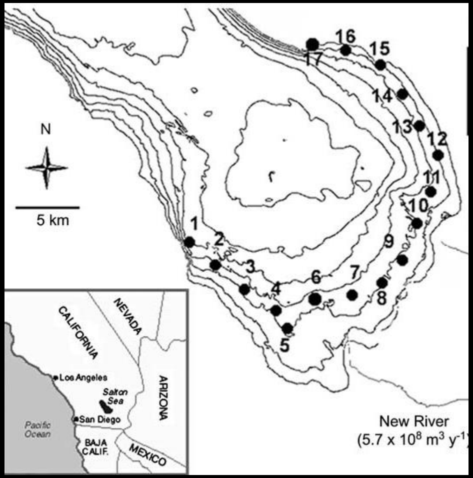
They answer : *"We can't. And should not try"*.

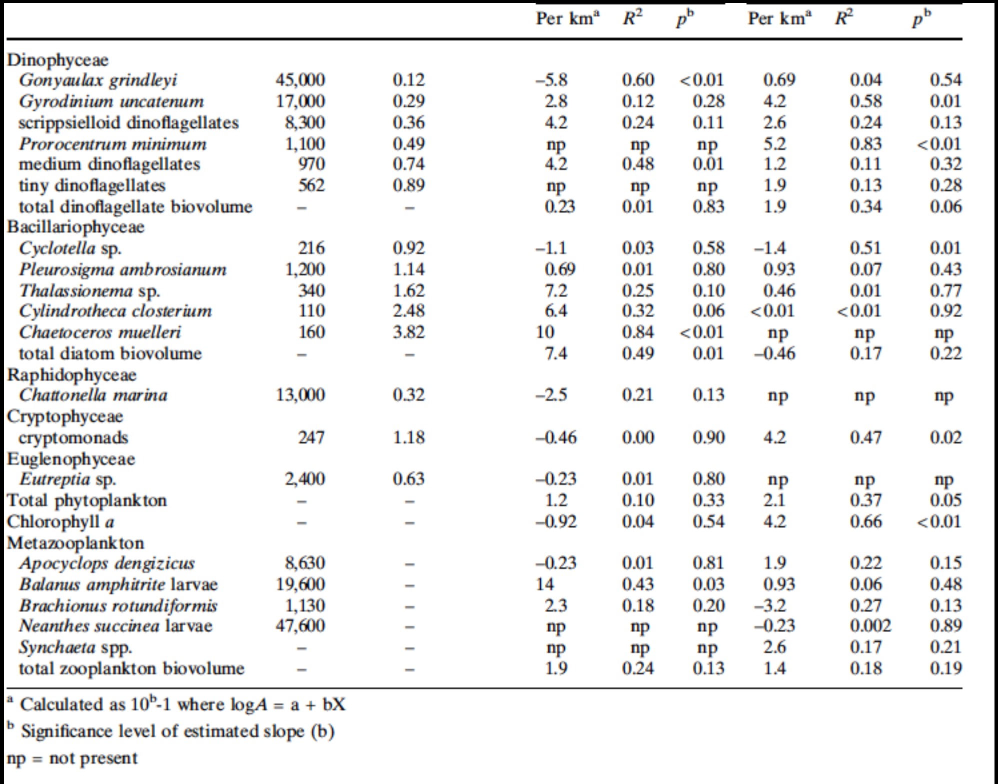
Recommend :

"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 ≤ 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 is the current status in too many branches of science:

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

Leaders such as Nature, NEJM, are not excluded

Hence it remains a silent killer of replicability even when their number is between a handful and a thousand

Summing up for evident selective inference

There is well developed theory and flexible practice to address evident selective inference, in not too power consuming way.

It may help you calibrate your 5σ

Adjusting for selection in estimation and confidence intervals is rarely practiced, even where done for testing, leading to dwindling results upon replication.

Collaborative research with many

Ruth Heller, Dani Yekutieli, Ilan Golani, Neri Kafkafi,	*TAU*
Marina Bogomolov,	*Technion*
Chiara Sabatti,	*Stanford*
Jonathan Rosenblatt	*Ben Gurion*
Philip Stark, Will Fithian	*Berkeley*
Asaf Weinstein	Carnegie Mellon
Yotam Hechtlinger	Carnegie Mellon
Christine Peterson,	Anderson MC
Iman Jaljuli,, Meir Amit, Yoav Zeevi	TAU

Supported by ERC grant: Practical Statistical Approaches to Replicability Problems in Life Sciences (PSARPS) and NIH-US Israel BSF grant

Thanks!

www.replicability.tau.ac.il

1888 1999

The industrialization of the scientific process

1950 2010

Defending the p-value

- It’s the first defense line against being fooled by randomness – needs minimal modeling assumptions
- Threshold for decision (selection) –

and selection is essential in modern science

likelihood ratio, posterior odds,…, are all practically subject to selection at a (sometimes) arbitrary threshold

Defending the p-value

- The meaning of p-value is shared across fields of science (like effect size)
- In some emerging branches of science it’s the only way to compare across conditions: GWAS, fMRI, Brain Networks, and here.

But it should not be allowed to be misused -

as any other method should not

So what did we have?

I hope I managed to convey

- The importance of offering rigorous but more lenient methods
- On-the-average-over-the selected in Hierarchical inference
- The challenges in addressing selective inference in database management

- Selective inference is generally not addressed and not well recognized as part of the replicability problem

The status: Bayesian statistics

Many Bayesian statisticians ignore the issue

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*.

The status: Bayesian statistics

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 μ_i~N(0,0.5^2); y_i~N(μ_i ,1); i=1,2,…,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 …

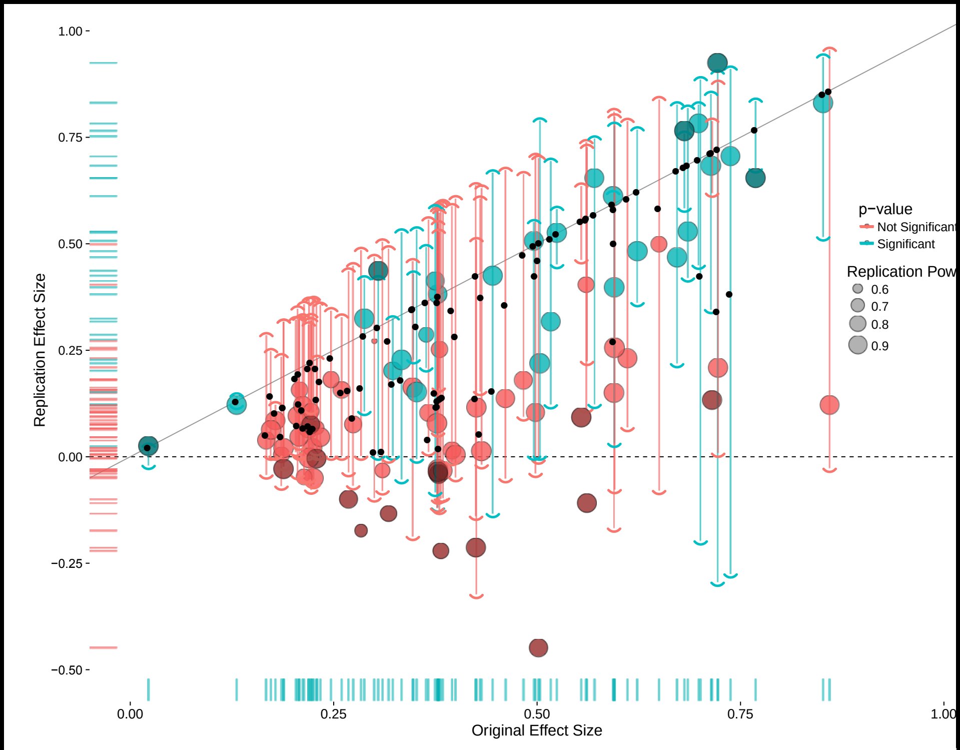
Purely Bayes model where selection should be addressed

Yekutieli et al ’13

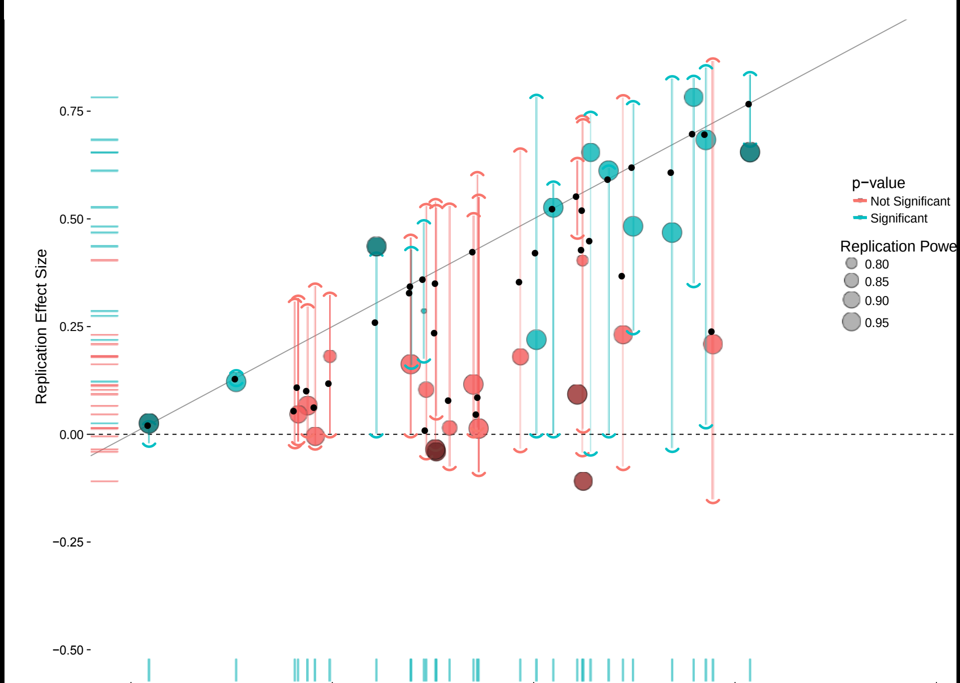
Thresholding of posterior odds using BH

Lowering the 0.05 threshold will not help

Thresholding at p-value ≤ 0.05



Thresholding at p-value ≤ 0.005



What can help?

6. Addressing the relevant variability

Mouse phenotyping example: opposite single lab results

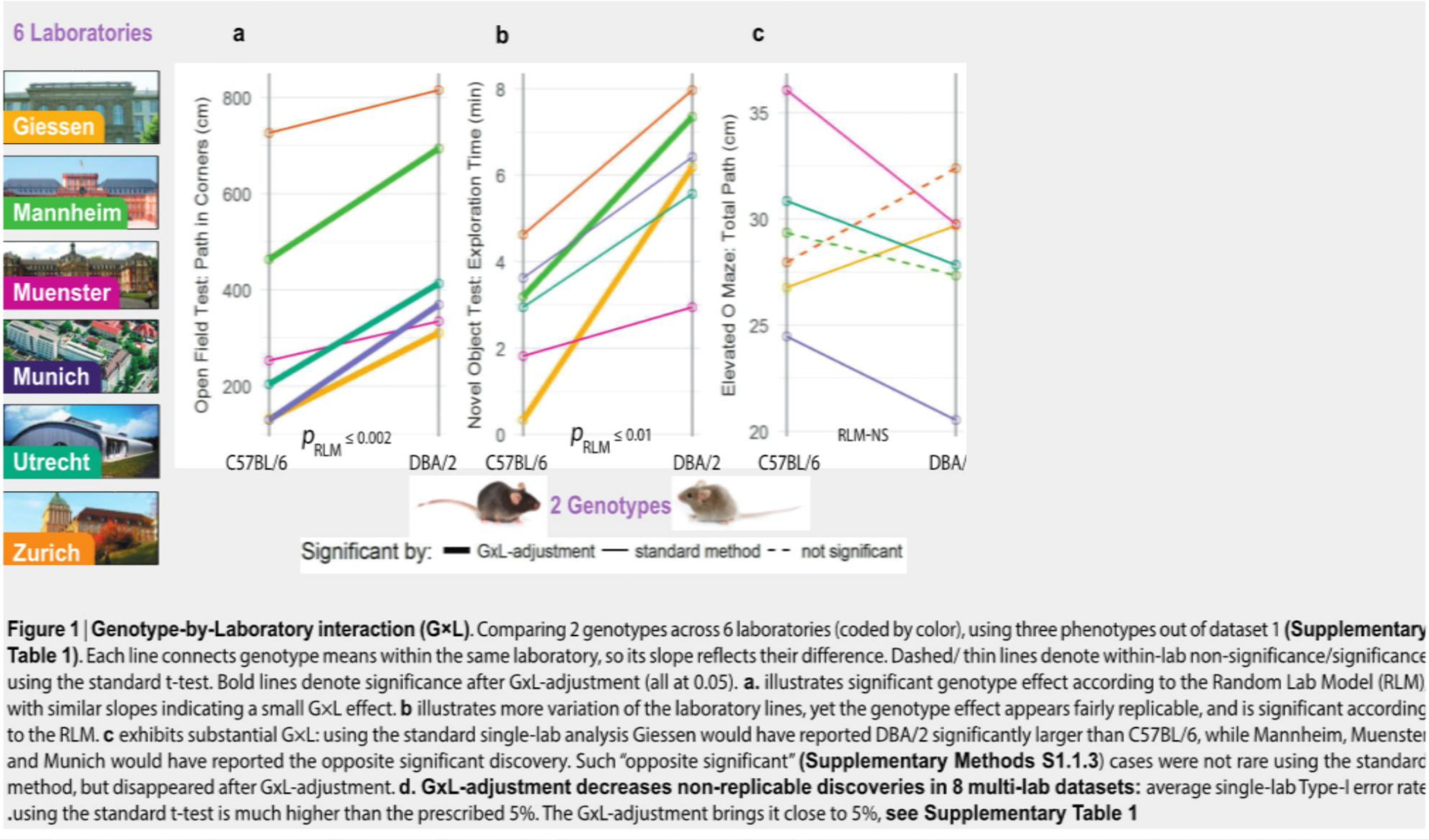


Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 **(Supplementary Table 1)**. Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/ thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). **a.** illustrates significant genotype effect according to the Random Lab Model (RLM) with similar slopes indicating a small GxL effect. **b** illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. **c** exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such "opposite significant" **(Supplementary Methods S1.1.3)** cases were not rare using the standard method, but disappeared after GxL-adjustment. **d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets:** average single-lab Type-I error rate .using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, **see Supplementary Table 1**

Kafkafi et al ('17 Nature Methods)

GxL interaction is “a fact of life”

Genotype-by-Lab effect for a genotype in a new lab is not known; but If its variability σ^2_{GxL}can be estimated, use

$$\frac{Mean(M_{G1}) - Mean(M_{G2})}{(\sigma^2_{Within}\ (1/n+1/n)+ 2\sigma^2_{GxL}\ )^{1/2}}$$

We call it GxL- adjustment

It’s the right “yardstick” against which genetic differences should be compared, when concerned with replicability.

6. Addressing the relevant variability

Mouse phenotyping example: opposite single lab results

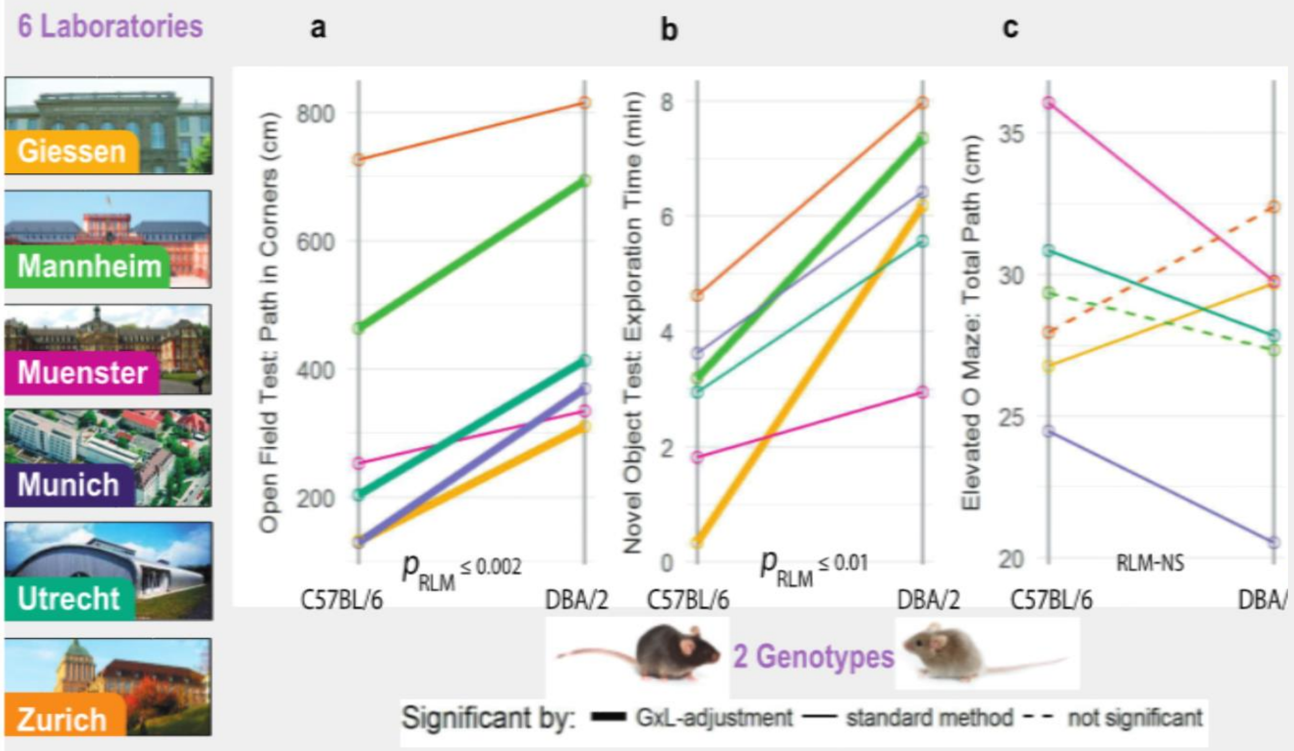


Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 **(Supplementary Table 1).** Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/ thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). **a.** illustrates significant genotype effect according to the Random Lab Model (RLM) with similar slopes indicating a small GxL effect. **b** illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. **c** exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such "opposite significant" **(Supplementary Methods S1.1.3)** cases were not rare using the standard method, but disappeared after GxL-adjustment. **d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets:** average single-lab Type-I error rate .using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, **see Supplementary Table 1**

Kafkafi et al ('17 Nature Methods)

Single-lab analyses in all known replication studies

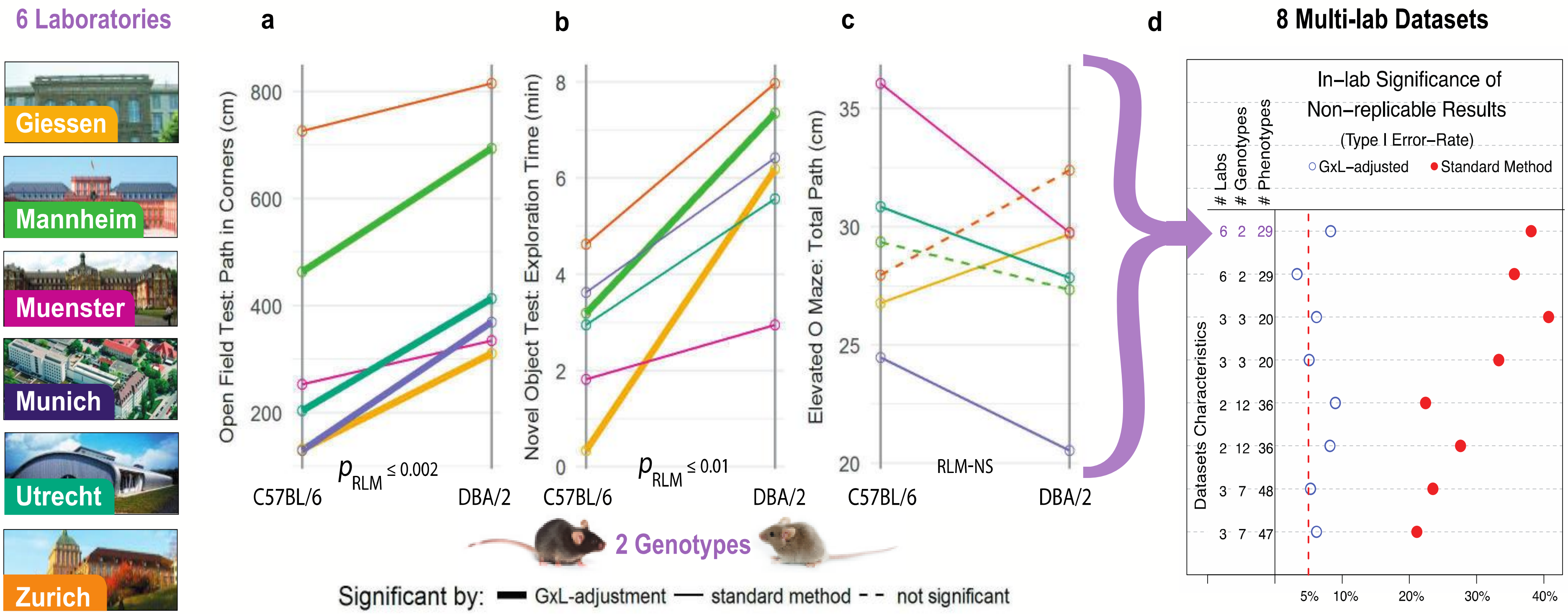


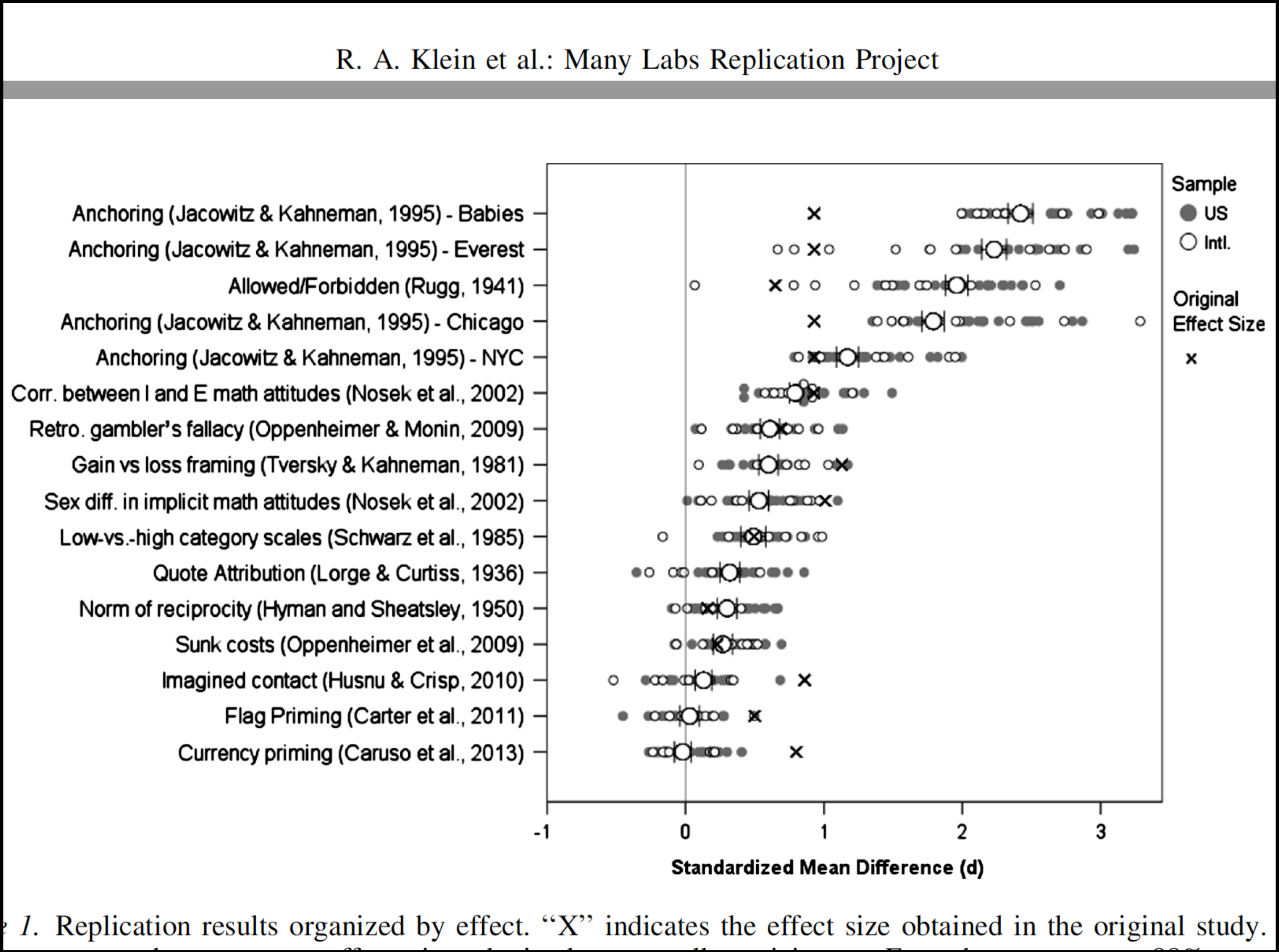
Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 **(Supplementary Table 1).** Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/ thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). **a.** illustrates significant genotype effect according to the Random Lab Model (RLM), with similar slopes indicating a small GxL effect. **b** illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. **c** exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such "opposite significant" **(Supplementary Methods S1.1.3**) cases were not rare using the standard method, but disappeared after GxL-adjustment. **d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets:** average single-lab Type-I error rate .using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, **see Supplementary Table 1**

Utilizing large database to get σ_{GxL}

1. Use available public database of mice phenotyping results (e.g. International Mouse Phenotyping Consor.) to estimate the interaction variability from the database (not statistical challenges free)
2. Scientists conducting experiments in their lab get an estimate of the relevant GxL variability
3. By enriching the database with their results future estimates will be improved

“Replicability Adjuster” Implemented at JAX Labs Bar Harbor

Kafkafi et al (Nature Methods ‘17)



R. A. Klein et al.: Many Labs Replication Project

1. Replication results organized by effect. “X” indicates the effect size obtained in the original study.

From the example to generality

Choosing the relevant level of variability is critical in order to increase replicability, for any inferential procedure: tests, confidence intervals, and estimates, Bayesian or frequentist.

It is a matter of precision vs generalizability

An old conflict

Unlike selective inference it has not become more severe

From the example to generality

Many small studies are better than single large one even if each is underpowered!

Clinical research: Multiple centers with Center by Treatment interaction . E.g. the Cochran reviews.

Educational research: Districts, schools, classes

Group Jackknife with groups reflecting relevant variability

In summary

Selective inference should be addressed

Getting rid of p-values, p<.05 or other bright-lines

Results in hiding them and worsens the problem

The relevant level of variability should be addressed

Still is merely Enhancing replicability

It is essential to increase the confidence of the public in the scientific method, and decrease waste of money and efforts,

But

Replicability cannot reliably be assessed without

actual replicability efforts by others

Replicating others' work as a way of life

Reproducibility projects are not sustainable.

Neither are publishing many papers with negative results only. Instead

- Every research proposal and paper should have a replicability-check component of a result, considered by the authors important for their proposed research.

Replicating others’ work as a way of life

- Granting agencies will support, but also review, the proposal for replication.
- It will be registered with Open Science Framework and its likes
- Its result will be reported whatever the outcome is, in the extended-abstract/main-body in 1-2 searchable sentences.

Replicating others' work as a way of life

- 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.

Replicating others' work as a way of life

Many weak studies can support not only the meta-analysis but a stronger statement of at least u out of m studies having consistent direction of effect using $r_{u|n}$ *: the smallest level at which the relevant null can be rejected .*

$r_{2|2}$ *is the simplest such statement which can be used to to assess Fisher's replicability.*

Heller Yekutieli, Bogomolov YB, Sofer, Wang, Jaljuli

Replicating others’ work as a way of life

- The authors of a replicated study will receive special recognition for having published a result considered important enough by others to invest the effort to replicate it. Unlike (see also [7]-[16])
- Involvement of researchers, granting agencies, publishers, academic leaders and of professional societies is needed.

Outline

1. The reproducibility and replicability crisis
2. The misguided attack
3. Selective inference
4. Addressing selective inference
5. The status of addressing evident selective inference
6. Addressing the relevant variability
7. Replication as a way of life in scientific work

5. The status of addressing evident selective inference

Bayesian statistics

Clinical trials

- for drug registration
- Other pre-registration
- Old and New NEJM

Large scale studies

Experimental Psychology

Open Science Framework

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

- Endpoints for efficacy primary and secondary
- Simultaneous over all primary endpoints.

If no primary endpoint shows statistical significance -

the study fails.

Hence their number is kept small.

Secondary endpoints? Safety ?

Fuels much statistical research in this area

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≤ 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 (in 84% a single one)

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

Back to Netalizumab case

Recall 50 secondary endpoints:

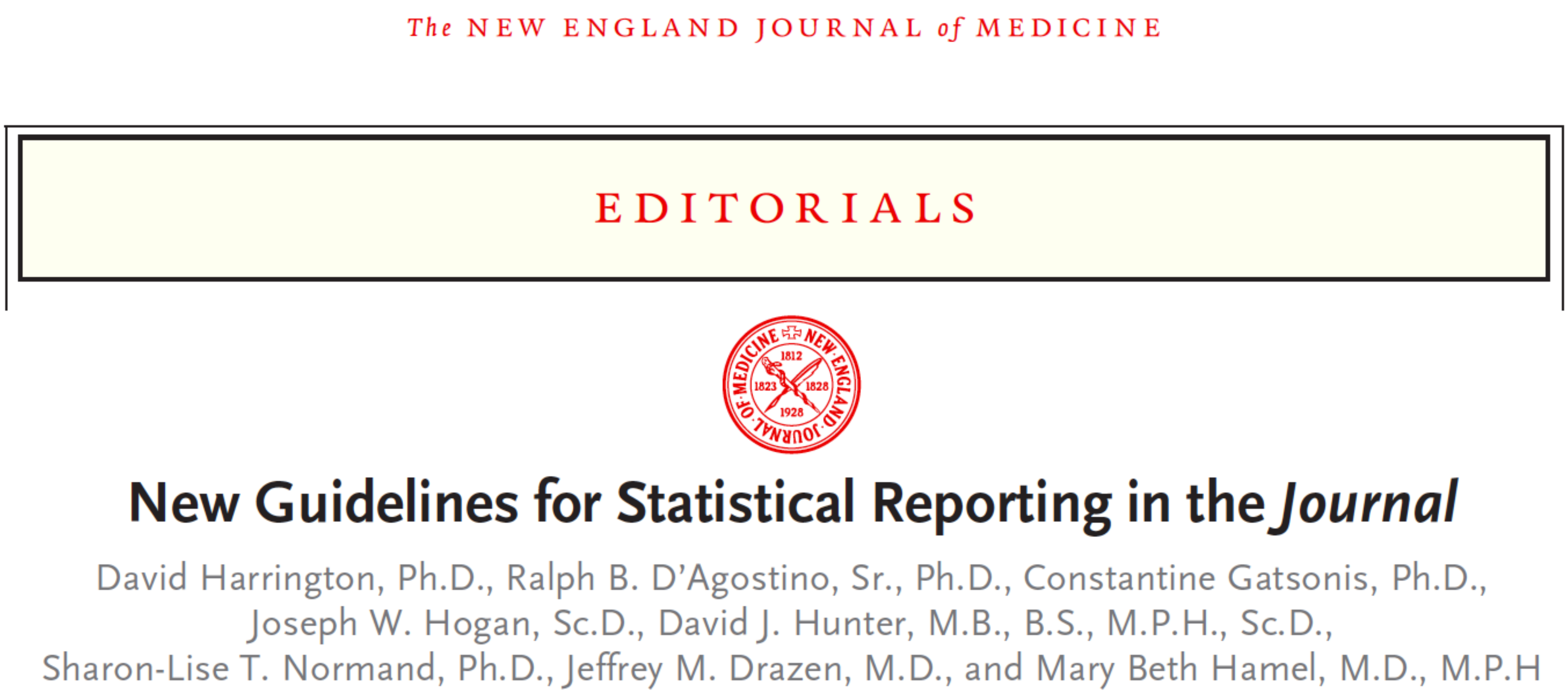
$$p_{S_{(1)}} = 4.76 \cdot 10^{-5}, p_{S_{(2)}} = 6.29 \cdot 10^{-5}, p_{S_{(3)}} = 1.44 \cdot 10^{-4}, \ldots, p_{S_{(50)}} = .992$$

Simes p-value for intersection of the secondaries

0.00157 $< 1/2 *$ .05

Pprimary 0.533 $> 1/2 *$ .05

12 secondary p-values $\leq 0.05 * 1/2 * 12/50$ rejected by Hierarchical BH while controlling the error rate (reporting adjusted p-values multiplied by half.) Study still as success.



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. “

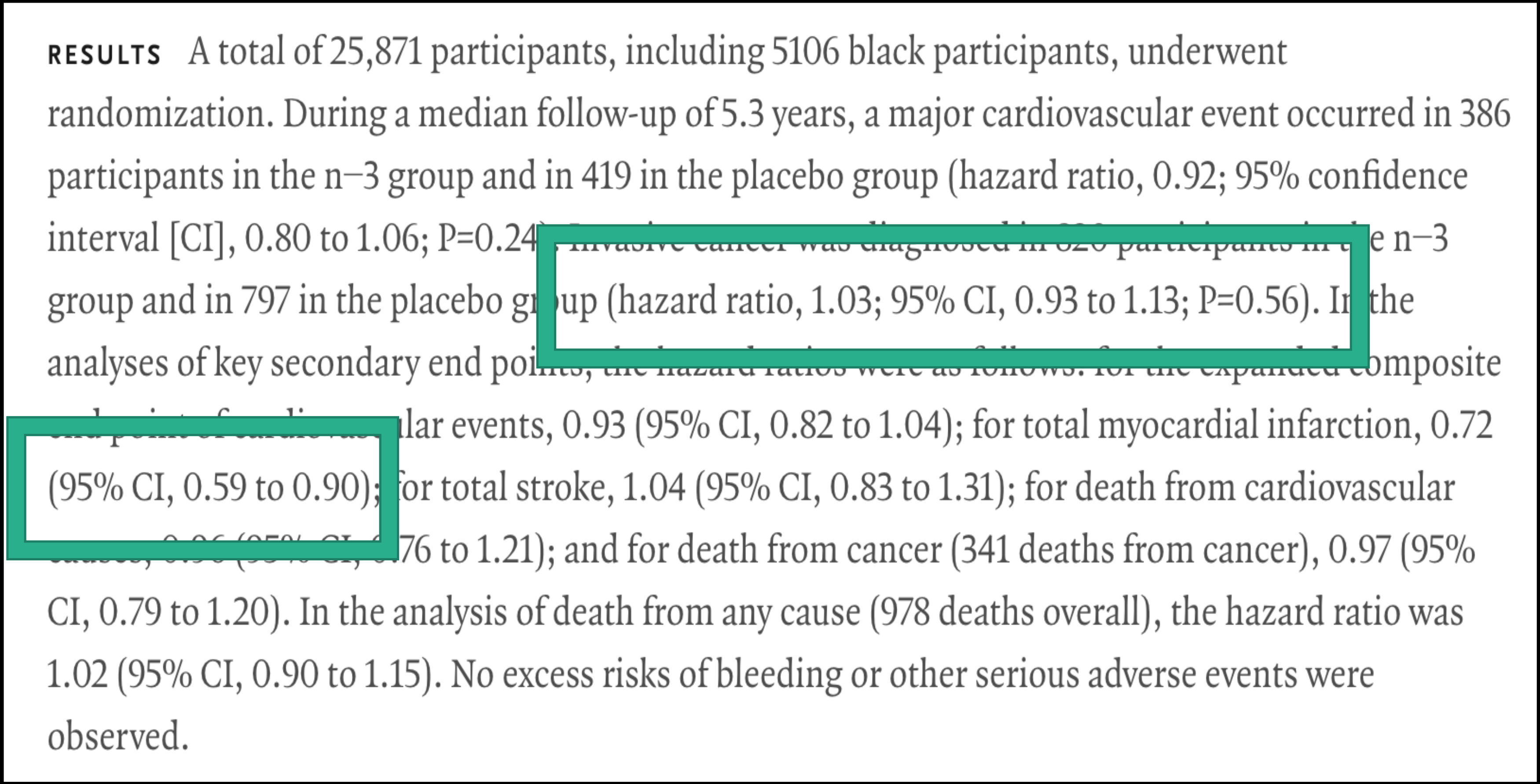
NEJM editorial July 2019 discussion

“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;

NEJM editorial July 2019 discussion

“However, the article reported only the hazard ratios and confidence intervals for the intervention effects for those secondary outcomes, consistent with recently implemented Journal guidelines limiting the use of P values for secondary and other comparisons.”

NEJM editorial July 2019 discussion



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 820 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.

The status: large scale studies (genomics, fMRI)

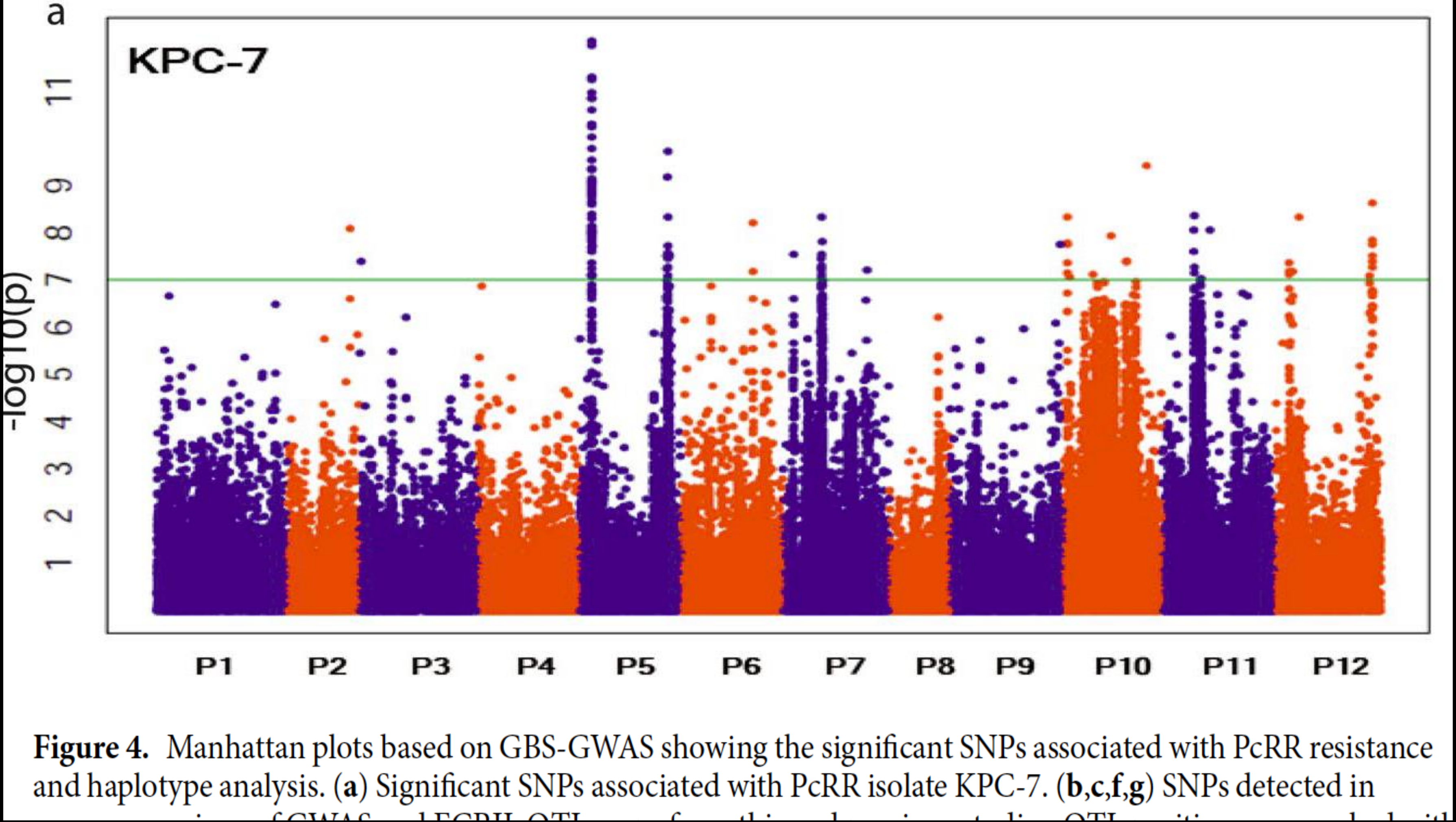
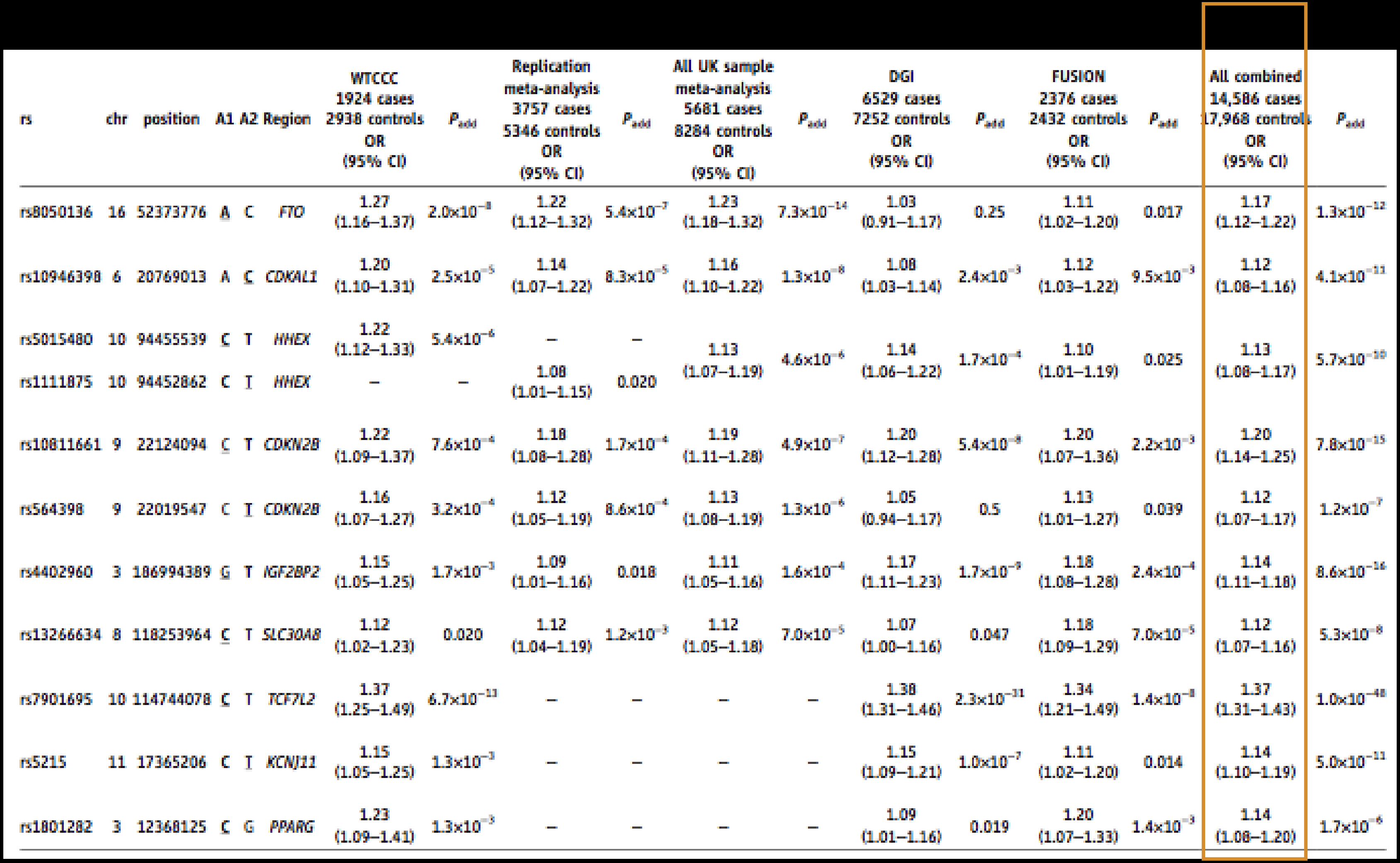


Figure 4. Manhattan plots based on GBS-GWAS showing the significant SNPs associated with PcRR resistance and haplotype analysis. (**a**) Significant SNPs associated with PcRR isolate KPC-7. (**b**,**c**,**f**,**g**) SNPs detected in

The bright line is actually green

The status: large scale studies (genomics, fMRI)



rs	chr	position	A1	A2	Region	WTCCC 1924 cases 2938 controls OR (95% CI)	P_{add}	Replication meta-analysis 3757 cases 5346 controls OR (95% CI)	P_{add}	All UK sample meta-analysis 5681 cases 8284 controls OR (95% CI)	P_{add}	DGI 6529 cases 7252 controls OR (95% CI)	P_{add}	FUSION 2376 cases 2432 controls OR (95% CI)	P_{add}	All combined 14,586 cases 17,968 controls OR (95% CI)	P_{add}
rs8050136	16	52373776	A	C	*FTO*	1.27 (1.16–1.37)	2.0×10^{-8}	1.22 (1.12–1.32)	5.4×10^{-7}	1.23 (1.18–1.32)	7.3×10^{-14}	1.03 (0.91–1.17)	0.25	1.11 (1.02–1.20)	0.017	1.17 (1.12–1.22)	1.3×10^{-12}
rs10946398	6	20769013	A	C	*CDKAL1*	1.20 (1.10–1.31)	2.5×10^{-5}	1.14 (1.07–1.22)	8.3×10^{-5}	1.16 (1.10–1.22)	1.3×10^{-8}	1.08 (1.03–1.14)	2.4×10^{-3}	1.12 (1.03–1.22)	9.5×10^{-3}	1.12 (1.08–1.16)	4.1×10^{-11}
rs5015480	10	94455539	C	T	*HHEX*	1.22 (1.12–1.33)	5.4×10^{-6}	–	–	1.13 (1.07–1.19)	4.6×10^{-6}	1.14 (1.06–1.22)	1.7×10^{-4}	1.10 (1.01–1.19)	0.025	1.13 (1.08–1.17)	5.7×10^{-10}
rs1111875	10	94452862	C	T	*HHEX*	–	–	1.08 (1.01–1.15)	0.020								
rs10811661	9	22124094	C	T	*CDKN2B*	1.22 (1.09–1.37)	7.6×10^{-4}	1.18 (1.08–1.28)	1.7×10^{-4}	1.19 (1.11–1.28)	4.9×10^{-7}	1.20 (1.12–1.28)	5.4×10^{-8}	1.20 (1.07–1.36)	2.2×10^{-3}	1.20 (1.14–1.25)	7.8×10^{-15}
rs564398	9	22019547	C	T	*CDKN2B*	1.16 (1.07–1.27)	3.2×10^{-4}	1.12 (1.05–1.19)	8.6×10^{-4}	1.13 (1.08–1.19)	1.3×10^{-6}	1.05 (0.94–1.17)	0.5	1.13 (1.01–1.27)	0.039	1.12 (1.07–1.17)	1.2×10^{-7}
rs4402960	3	186994389	G	T	*IGF2BP2*	1.15 (1.05–1.25)	1.7×10^{-3}	1.09 (1.01–1.16)	0.018	1.11 (1.05–1.16)	1.6×10^{-4}	1.17 (1.11–1.23)	1.7×10^{-9}	1.18 (1.08–1.28)	2.4×10^{-4}	1.14 (1.11–1.18)	8.6×10^{-16}
rs13266634	8	118253964	C	T	*SLC30A8*	1.12 (1.02–1.23)	0.020	1.12 (1.04–1.19)	1.2×10^{-3}	1.12 (1.05–1.18)	7.0×10^{-5}	1.07 (1.00–1.16)	0.047	1.18 (1.09–1.29)	7.0×10^{-5}	1.12 (1.07–1.16)	5.3×10^{-8}
rs7901695	10	114744078	C	T	*TCF7L2*	1.37 (1.25–1.49)	6.7×10^{-13}	–	–	–	–	1.38 (1.31–1.46)	2.3×10^{-31}	1.34 (1.21–1.49)	1.4×10^{-8}	1.37 (1.31–1.43)	1.0×10^{-48}
rs5215	11	17365206	C	T	*KCNJ11*	1.15 (1.05–1.25)	1.3×10^{-3}	–	–	–	–	1.15 (1.09–1.21)	1.0×10^{-7}	1.11 (1.02–1.20)	0.014	1.14 (1.10–1.19)	5.0×10^{-11}
rs1801282	3	12368125	C	G	*PPARG*	1.23 (1.09–1.41)	1.3×10^{-3}	–	–	–	–	1.09 (1.01–1.16)	0.019	1.20 (1.07–1.33)	1.4×10^{-3}	1.14 (1.08–1.20)	1.7×10^{-6}

GWAS for type II Diabetics:11 selected by the table out of ~400,000

Confidence intervals are not adjusted

The status: Experimental psychology

Our analysis of the 100 in the Psychology reproducibility project:

Zeevi, Meir, Estachenko,

# of inferences per study (4-700, average 72);

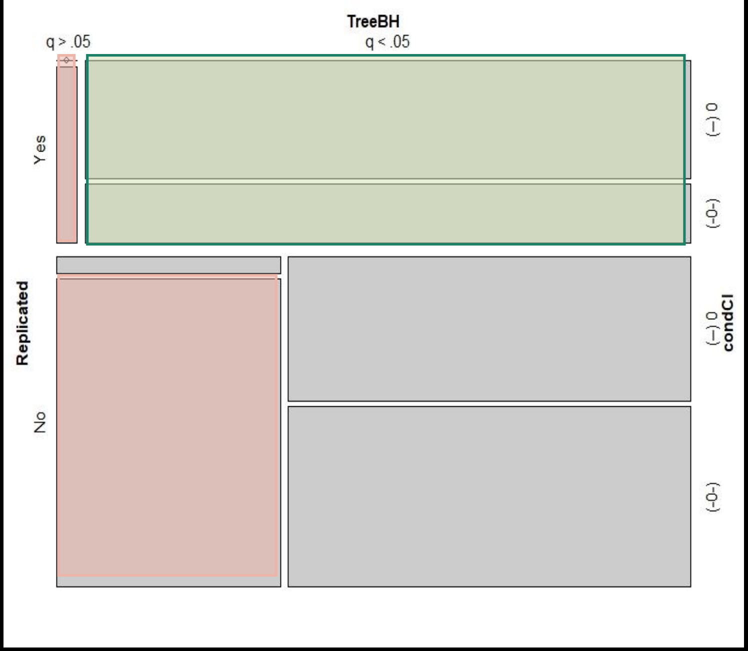
Only 11 (very very partially) addressed selection

8 had reproducibility error $p \leq .05$ was $p > .05$

The status: experimental psychology

Tree BH testing

q>.05 q≤.05



Yes

Replicated

No

No

Yes

Conditional CI covers 0

No

Yes

FDR control: Of 22 with q>.05, 21 were un-replicated

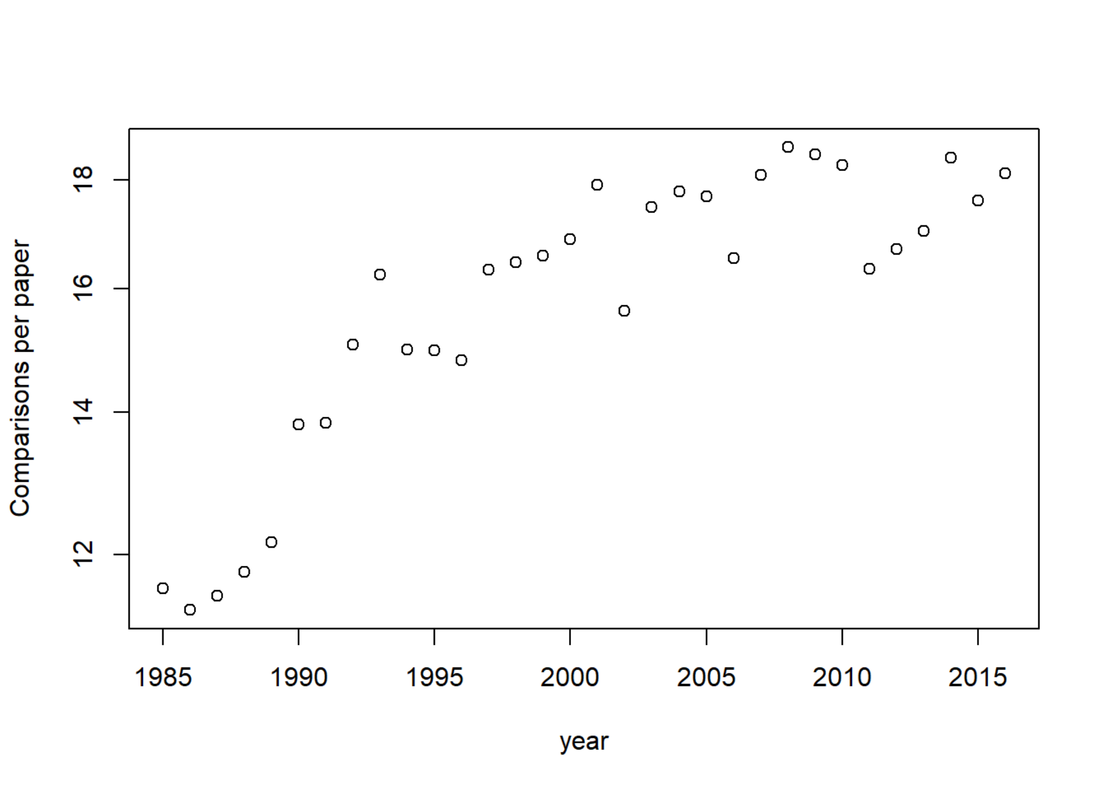
False Discovery proportion 30/66=.45 (.40)

instead of 57/87= .66

Power 30/31= .96

The status: Experimental psychology

Number of inferences only in the form ‘p=‘ or ‘p<‘ or ‘p>’ in a paper averaging over all papers per year



Discussion and call for action (e.g. Glickman et al ’14)- with no impact

The status: Open Science Framework

Leaders in their efforts to offer tools for pre-registered and transparent research

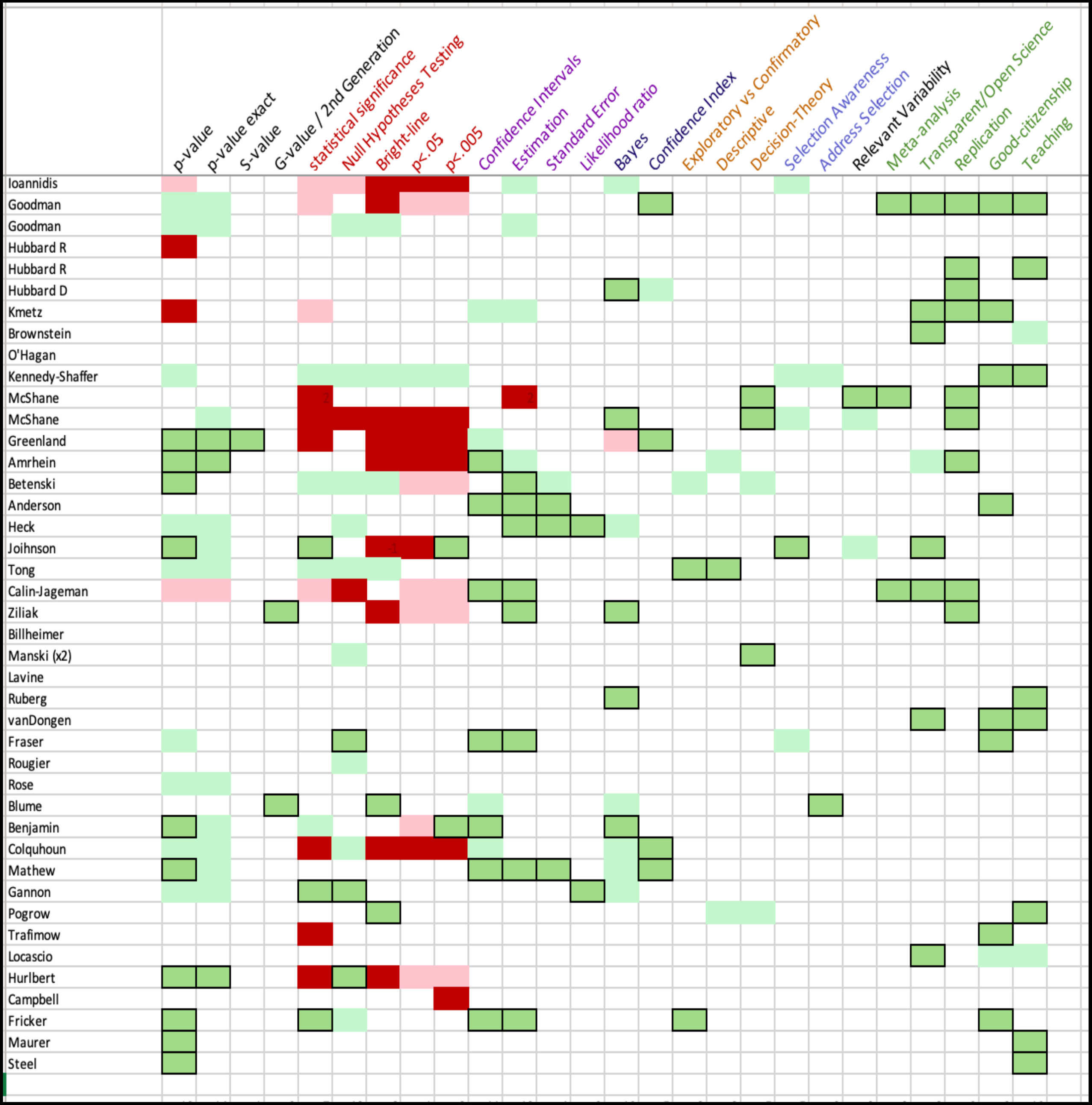
“If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?”

The status: Open Science Framework

Exemplary study: compares two methods on 29 items.

The item scores’ sum is the endpoint in the main analysis.

”With respect to the follow-up analyses, we will also maintain an alpha of 1/20 for each of the analyses. We will not use any other inference criteria for these follow-up analyses.”



	p-value	p-value exact	S-value	G-value / 2nd Generation	statistical significance	Null Hypotheses Testing	Bright-line	p<.05	p<.005	Confidence Intervals	Estimation	Standard Error	Likelihood ratio	Bayes	Confidence Index	Exploratory vs Confirmatory	Descriptive	Decision-Theory	Selection Awareness	Address Selection	Relevant Variability	Meta-analysis	Transparent/Open Science	Replication	Good-citizenship	Teaching
Ioannidis	🩷				🩷	🩷	🟥	🟥	🟥		🟢			🟢					🟢							
Goodman	🟢	🟢			🩷		🟥	🩷	🩷						🟩							🟩	🟩	🟩	🟩	🟩
Goodman	🟢	🟢				🟢	🟢				🟢															
Hubbard R	🟥																									
Hubbard R																								🟩		🟩
Hubbard D														🟩	🟢									🟩		
Kmetz	🟥				🩷					🟢	🟢												🟩	🟩	🟩	
Brownstein																							🟩			🟢
O'Hagan																										
Kennedy-Shaffer	🟢				🟢	🟢	🟢	🟢	🟢										🟢	🟢					🟩	🟩
McShane					🟥 2						🟥 2							🟩			🟩	🟩		🟩		
McShane		🟢			🟥	🟥	🟥	🟥	🟥					🟩				🟩	🟢		🟢			🟩		
Greenland	🟩	🟩	🟩		🟥		🟥	🟥	🟥	🟢				🩷	🟩											
Amrhein	🟩	🟩					🟥	🟥	🟥	🟩	🟢						🟢						🟢	🟩		
Betenski	🟩				🟢	🟢	🟢	🩷	🩷		🟩	🟢				🟢		🟢								
Anderson										🟩	🟩	🟩													🟩	
Heck	🟢	🟢				🟢					🟩	🟩	🟩	🟢												
Joihnson	🟩	🟢			🟩		🟥 -1	🟥	🟩										🟩		🟢		🟩			
Tong	🟢	🟢			🟢	🟢	🟢									🟩	🟩									
Calin-Jageman	🩷	🩷			🩷	🟥		🩷	🩷	🟩	🟩											🟩	🟩	🟩		
Ziliak				🟩			🟥	🩷	🩷		🟩			🟩										🟩		
Billheimer																										
Manski (x2)						🟢												🟩								
Lavine																										
Ruberg														🟩												🟩
vanDongen																							🟩		🟩	🟩
Fraser	🟢					🟩				🟩	🟩								🟢						🟩	
Rougier						🟢																				
Rose	🟢	🟢																								
Blume				🟩			🟩			🟢				🟢						🟩						
Benjamin	🟩	🟢			🟢			🩷	🟩	🟩				🟩												
Colquhoun	🟢	🟢			🟥	🟢	🟥	🟥	🟥	🟢				🟢	🟩											
Mathew	🟩	🟢								🟩	🟩	🟩		🟢	🟩											
Gannon	🟢	🟢			🟩	🟩							🟩	🟢												
Pogrow							🟩										🟢	🟢								🟩
Trafimow					🟥																				🟩	
Locascio																							🟩		🟢	🟢
Hurlbert	🟩	🟩			🟥	🟩	🟥	🩷	🩷																	
Campbell									🟥																	
Fricker	🟩				🟩	🟢				🟩	🟩					🟩									🟩	
Maurer	🟩																									🟩
Steel	🟩																									🟩

Moving to a world beyond 'p<.05'

Summary of the 43 papers by

Wasseerstein, Schirm & Lazar

- Don't use p<.05
- Don't say "statistically significant"
- or use any bright-line rule

There are too many Do's

Accept Uncertainty.

Be Thoughtful, Open

And Modest

A Replicability Crisis

Has turned into

A Statistical Crisis

Karen Kafadar’s responding to the TAS (Amstat News June)

Nonstatisticians …may be confused about what to do. Worse, “by breaking free from the bonds of statistical significance” as the editors suggest and several authors urge, researchers may read the call to “abandon statistical significance” as “abandon statistical methods altogether.”

Karen Kafadar's responding to the TAS (Amstat News June)

"We should take responsibility for the situation in which we find ourselves today …

to ensure that our well-researched and theoretically sound statistical methodology is neither abused nor dismissed categorically."

What other approaches were mentioned?

Confidence intervals

Prediction intervals

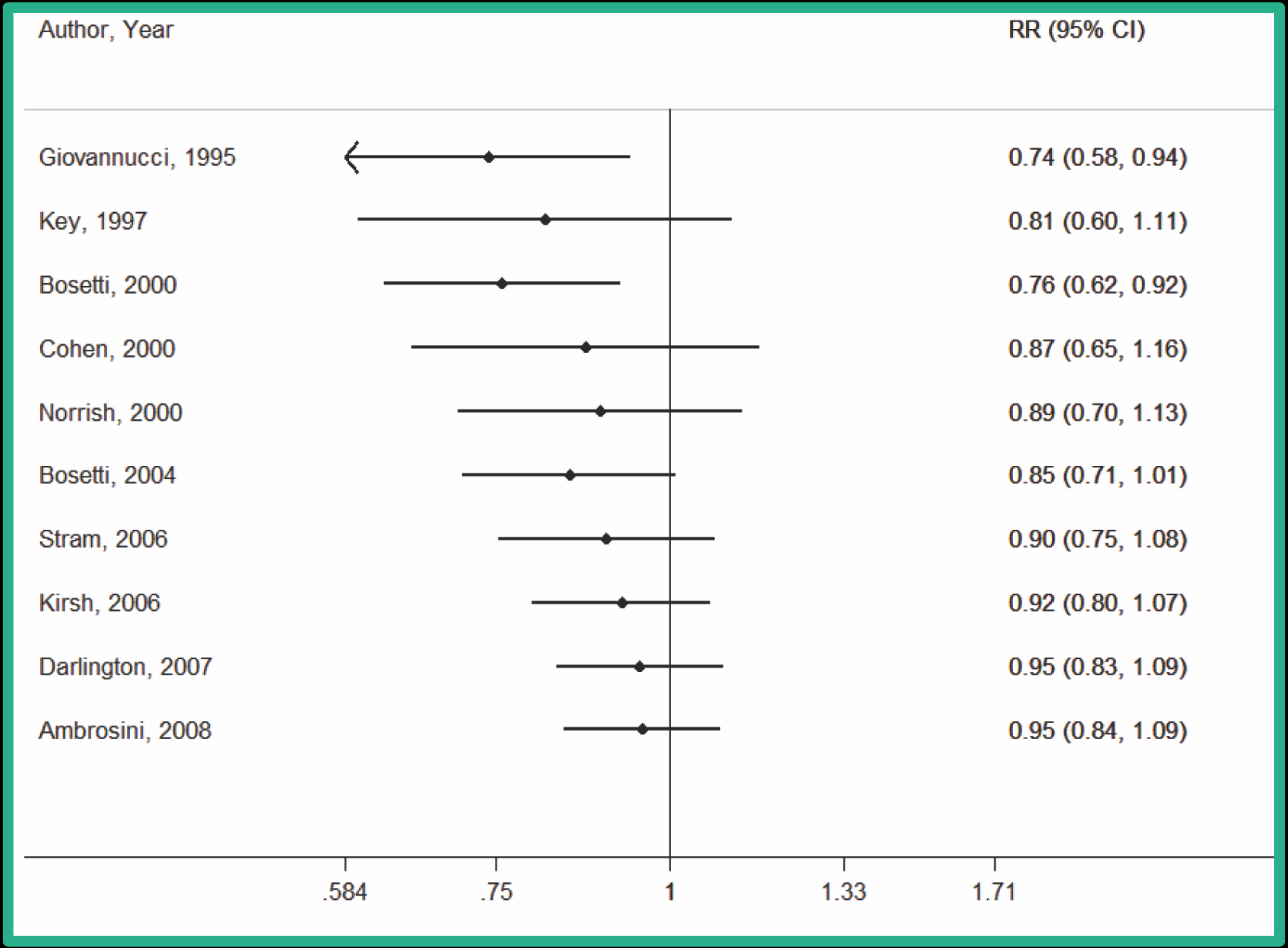
Estimation

Likelihood ratios

Bayesian methods

Bayes factor

Credibility intervals



"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)

Ban them!

Basic and Applied Social Psychology

Editorial by Trafimow & Marks Feb 24, 2015

"From now on, BASP is banning the NHSTP...prior to publication, authors will have to remove all vestiges of the NHSTP (p -values, t -values, F –values, statements about ''significant'' differences or lack thereof, and so on)."

Is it the p-values' fault?

A year long process started by American Statistical Association (ASA).

ASA Board's statement about p-values (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"

Newer solutions

Stepwise procedures that make use of observed p-values:

Let P_i be the observed p-value of the test for H_i

(4) Holm's procedure: **Always** α

Order the p-values $P_{(1)} \le P_{(2)} \le \dots P(k) \le P(k+1) \le \dots P_{(m)}$

Reject as long as $\le$ $\le$ $\le$ $>$ $>$

α/m $\alpha/(m-1)$ $\alpha/(m+1-k)$ $\alpha/(m-k)$ α

Until for the first time $P_{(k)} > \alpha/(m+1-k)$, *then stop*

(5) Hochberg's procedure: **Positive dependence and ind.**

Accept as long as

Until for the first time $P_{(k)} \le \alpha/(m+1-k)$, *then stop and reject all k*