The question of reproducibility in brain imaging genetics

Jean-Baptiste Poline
jbpoline@berkeley.edu

Henry Wheeler Jr. Brain Imaging Center,
Helen Wills Neuroscience Institute, UC Berkeley, CA

June 10, 2015

The problem of reproducibility for Imaging genetics

Jean-Baptiste Poline

Henry Wheeler Brain Imaging Center,
Helen Wills Neuroscience Institute,
University of California Berkeley

Reproducibility - preliminary remarks

- Reminding ourselves : Reproducibility is the backbone of scientific activity
- Reproducibility versus replicability
- Is there a problem ?
- Plan:
  - Evidence for the problem
  - Causes: especially power issues
  - What should we do

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"
- In social sciences and in psychology

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"
- In social sciences and in psychology
  - Reproducibility Project: Psychology (open science foundation)

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"
- In social sciences and in psychology
  - Reproducibility Project: Psychology (open science foundation)
  - Simmons, et al. "... Undisclosed Flexibility ... Allows Presenting Anything as Significant." 2011.

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"
- In social sciences and in psychology
  - Reproducibility Project: Psychology (open science foundation)
  - Simmons, et al. "... Undisclosed Flexibility ... Allows Presenting Anything as Significant." 2011.
- In cognitive neuroscience

Reproducibility - evidence of the problem

- In general: Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
  - Revised standard for statistical evidence (PNAS 2013)
- In epidemiology
  - Ioannidis 2011: "The FP/FN Ratio in Epidemiologic Studies:"
- In social sciences and in psychology
  - Reproducibility Project: Psychology (open science foundation)
  - Simmons, et al. "... Undisclosed Flexibility ... Allows Presenting Anything as Significant." 2011.
- In cognitive neuroscience
  - Barch, Deanna M., and Tal Yarkoni. "Special Issue on Reliability and Replication in Cognitive and Affective Neuroscience Research." 2013.

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): “6 out of 53 key findings could not be replicated”
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007
  - Thirion et al., 2007: reproducibility of second level analyses

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007
  - Thirion et al., 2007: reproducibility of second level analyses
- In genetics

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007
  - Thirion et al., 2007: reproducibility of second level analyses
- In genetics
  - Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: ". . . results are largely null."

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007
  - Thirion et al., 2007: reproducibility of second level analyses
- In genetics
  - Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: ". . . results are largely null."
  - Many references and warning: eg:"Drinking from the fire hose . . . " by Hunter and Kraft, 2007.

Reproducibility - evidence of the problem

- Oncology Research:
  - Begley C.G. & Ellis L. Nature, (2012): "6 out of 53 key findings could not be replicated"
- In brain imaging
  - Reproducibility Issues in Multicentre MRI Studies, J. Jovicich
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007
  - Thirion et al., 2007: reproducibility of second level analyses
- In genetics
  - Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: ". . . results are largely null."
  - Many references and warning: eg:"Drinking from the fire hose . . . " by Hunter and Kraft, 2007.
- And in imaging genetics ?

Why do we have a problem?

- Things are getting complex
  - Data description, data size, computations, statistical methods
- Publication pressure is high
  - Cannot afford *not* to have a paper out of this data set - competitive research
- Mistakes are done
  - cf quite a few examples (R/L, Scripts errors (ADHD 1000FC), Siemens slice order, ...
  - but how many are *not* found ?
  - "The scientific method's central motivation is the ubiquity of error — the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error." *Donoho, 2009.*
- ***Power issues***

The power issue

- Ioannidis 2005: *"Why most research findings are false"*
- Button et al. 2013: *"Power failure"*
- Remember what is power
- What are the issues of low powered studies
- Tools to compute power
- What is our effect size?

The power issue

What is the effect ?

$$\mu = \bar{x_1} - \bar{x_2}$$

What is the standardized effect ? (eg Cohen's d)

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sigma} = \frac{\mu}{\sigma}$$

"Z" : Effect accounting for the sample size

$$Z = \frac{\mu}{\sigma/\sqrt{n}}$$

The power issue

What exactly is power ?

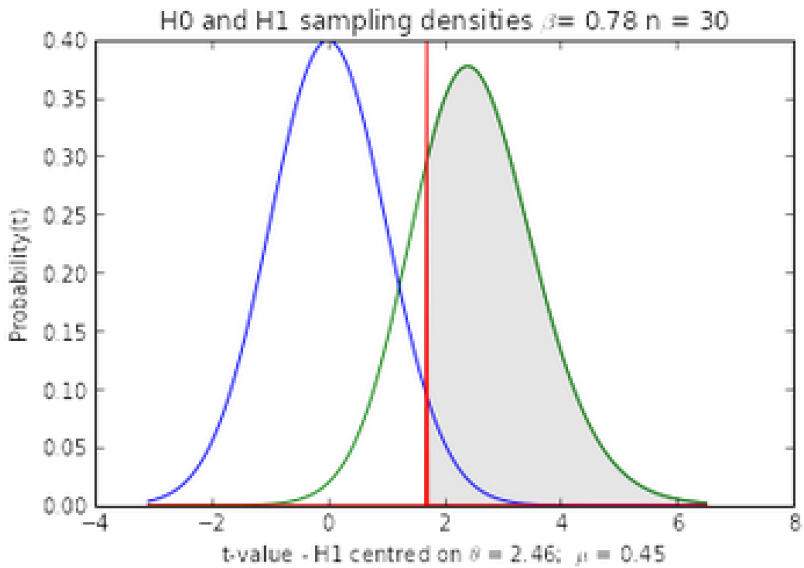


Figure 1: Power: $W = 1 - \beta$ Here W=77%

Cohen's d and relation with n :

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sigma} = \frac{\mu}{\sigma} \qquad Z = \frac{\mu\sqrt{n}}{\sigma} = d\sqrt{n}$$

The power issue

- Studies of low power have low probability of detecting an effect (indeed!)
- Studies of low power have low positive predictive value: $PPV = P(H1\,True|Detection)$
- Studies of low power are likely to show inflated effect size

The power issue

- $PPV = P(H1\,True|Detection) = \frac{W\,P_1}{\alpha\,P_0 + W\,P_1}$
- If we have 4/5 that H0 is true, and 1/5 that H1 true, with 30% power: PPV = 60%.

P1/P0 =0.25	power=0.10,	alpha=0.05	PPV=0.33
P1/P0 =0.25	power=0.30,	alpha=0.05	PPV=0.60
P1/P0 =0.25	power=0.50,	alpha=0.05	PPV=0.71
P1/P0 =0.25	power=0.70,	alpha=0.05	PPV=0.78

The power issue

What happens with more stringent α?

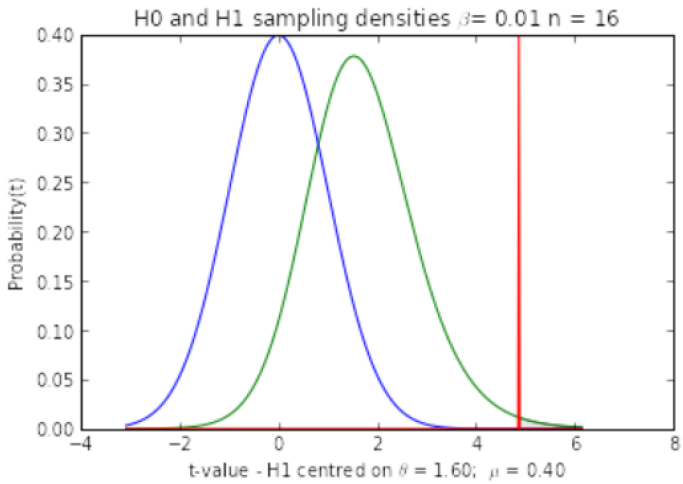


Figure 2: higher type I error threshold to account for MC

- effect on power: power goes down
- effect on PPV: PPV goes up
- effect on estimated effect size: size bias: goes up

The power issue

Studies of low power inflate the detected effect (2)

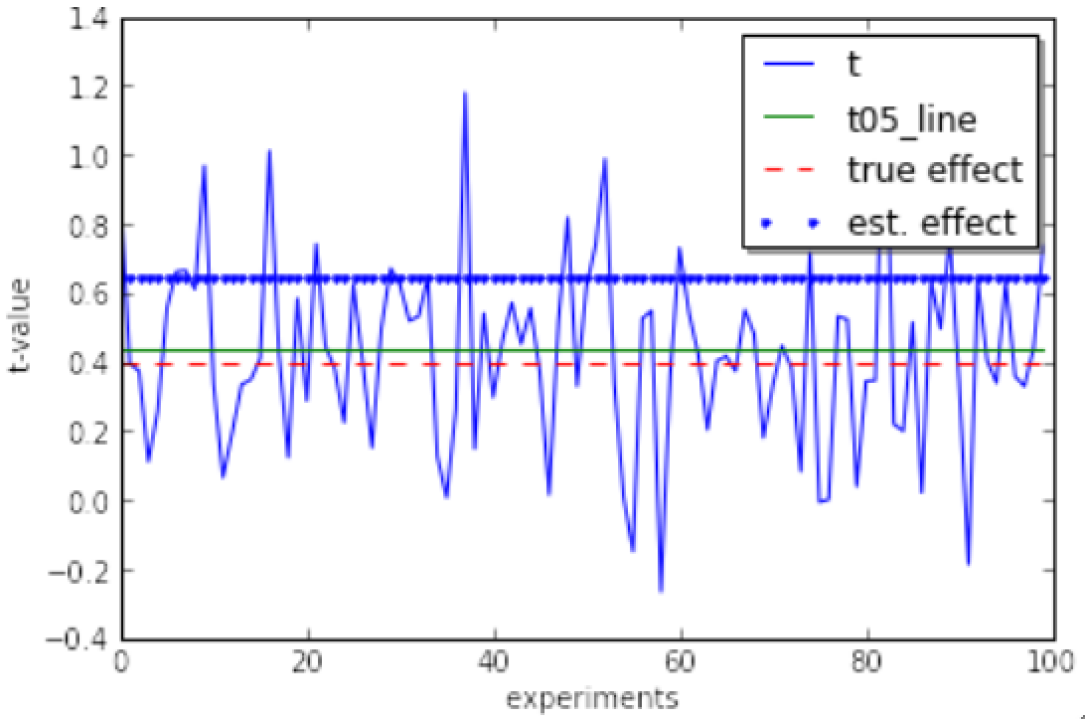


Figure 3: Repeating experiments: estimated effects are above t05 line, leading to a biased estimation compared to true simulated effect.

The power issue

Studies of low power inflate the detected effect (1)

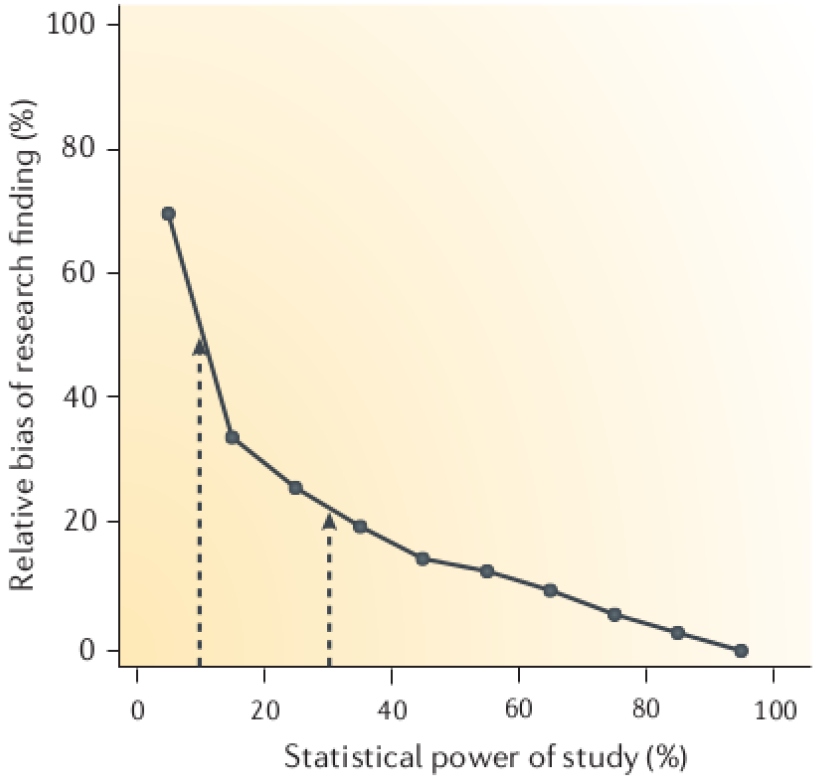


Figure 4: Button et al. NRN, 2013

The power issue

What is the estimated power in common meta analyses?

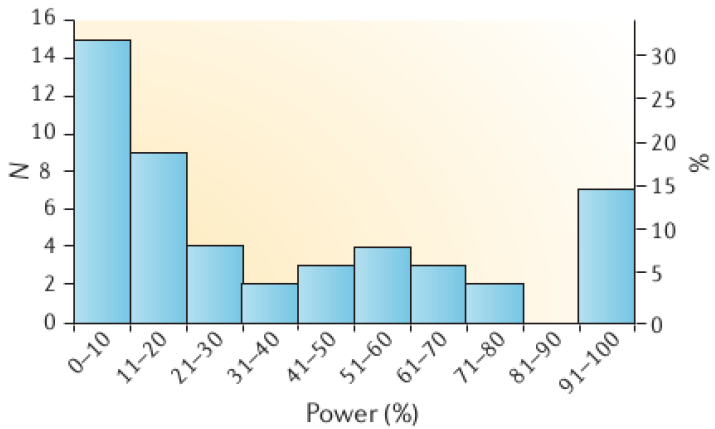


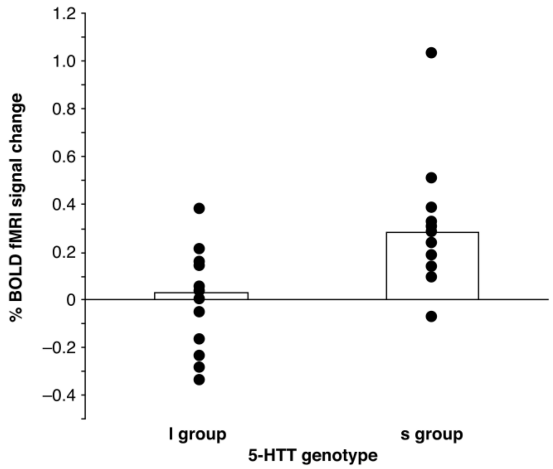
Figure 5: Button et al. NRN, 2013

What is specific to Imaging Genetics

- Combinaison of imaging and of genetics issues ("AND" problem)
- The combination of having to get very large number of subjects for GWAS and not being able to get them in imaging
- The multiple comparison issues
- The "trendiness" of the field
- The flexibility of analyses / exploration
- The capacity to "rationalize findings"
  - noise in brain images is always interpretable
  - genes are always interpretable

Computing effect size in imaging genetics (1)

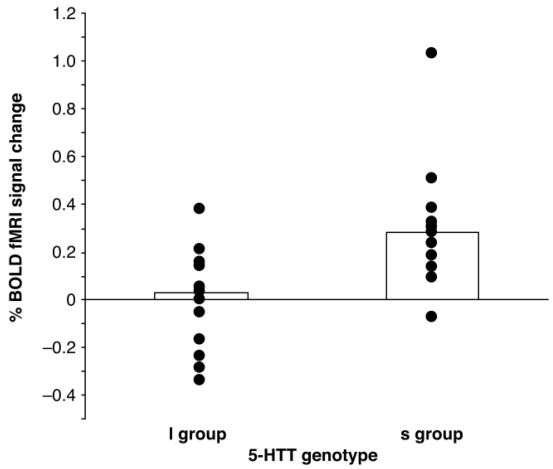
- Example of Hariri 2002: In Fig 3, Authors report
  $m_1 = .28, m_2 = .03, \mathrm{SDM}_1 = 0.08, \mathrm{SDM}_2 = 0.05, n_1 = n_2 = 14$



- What is the effect size ? Compute

Computing effect size in imaging genetics (1)

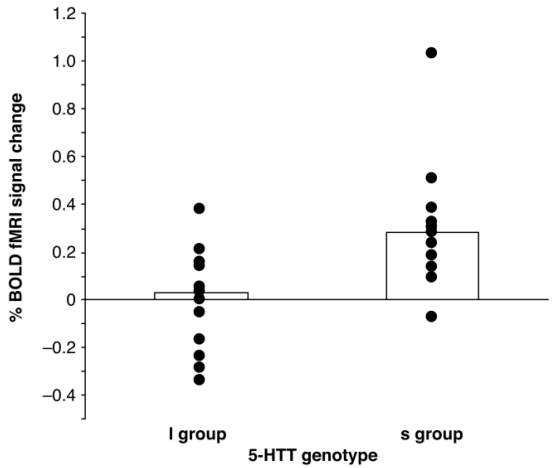
- Example of Hariri 2002: In Fig 3, Authors report
  $m_1 = .28, m_2 = .03, \mathrm{SDM}_1 = 0.08, \mathrm{SDM}_2 = 0.05, n_1 = n_2 = 14$



- What is the effect size ? Compute
  - $s_{1,2} = \sqrt{(14-1)}\mathrm{SDM}_{1,2}$, $d = \frac{m_1 - m_2}{s} = 1.05$

Computing effect size in imaging genetics (1)

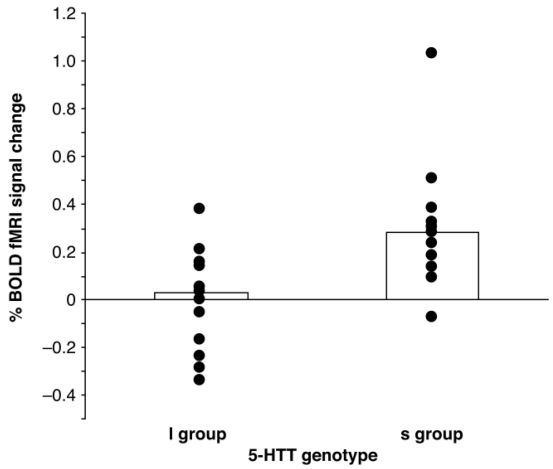
- Example of Hariri 2002: In Fig 3, Authors report
  $m_1 = .28, m_2 = .03, \mathrm{SDM}_1 = 0.08, \mathrm{SDM}_2 = 0.05, n_1 = n_2 = 14$



- What is the effect size ? Compute
  - $s_{1,2} = \sqrt{(14-1)\mathrm{SDM}_{1,2}}$, $d = \frac{m_1 - m_2}{s} = 1.05$
- What is the percentage of variance explained ?

Computing effect size in imaging genetics (1)

- Example of Hariri 2002: In Fig 3, Authors report
  $m_1 = .28, m_2 = .03, \mathrm{SDM}_1 = 0.08, \mathrm{SDM}_2 = 0.05, n_1 = n_2 = 14$



- What is the effect size ? Compute
  - $s_{1,2} = \sqrt{(14-1)\mathrm{SDM}_{1,2}}$, $d = \frac{m_1 - m_2}{s} = 1.05$
- What is the percentage of variance explained ?
  - $V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$

Computing effect size in imaging genetics (2)

- Example of Shen et al using the ADNI cohort: Association of SNPs and the amount of GM in the hippocampus.
- N = 733 subjects, considered a large study for imaging, but a very small one for genome wide association.
- only APOE gene confirmed, p = 6.63e-10: reaches GWAS significance level of 5.10-8
- Effect size for APOE ?
  - In [2]: n01.isf(6.63e-10) #- from p to Z value
  - Out[2]: 6.064
  - In [3]: n01.isf(6.63e-10)/sqrt(733) #- Correct for the number of subjects
  - Out[3]: 0.22

Effect size and reproducibility?

- Effect size in imaging genetics:
  - BDNF and hippocampal volume: genuine effect or winners curse? d=0.12, p=0.02, Molendijk (2012)
  - Stein et al, 2012: marker is associated with 0.58% of intracranial volume per risk allele
  - Flint 2014: Effect size of intermediate phenotype not much greater than others
  - For psychiatric diseases: mean OR is 1.15, QT: variance explained by 1 locus $<<$ 0.5%, 0.1-0.3% for protein or serum concentration
- Unlikely effect sizes

Effect size and reproducibility?

- Effect size in imaging genetics:
  - BDNF and hippocampal volume: genuine effect or winners curse? d=0.12, p=0.02, Molendijk (2012)
  - Stein et al, 2012: marker is associated with 0.58% of intracranial volume per risk allele
  - Flint 2014: Effect size of intermediate phenotype not much greater than others
  - For psychiatric diseases: mean OR is 1.15, QT: variance explained by 1 locus $<<$ 0.5%, 0.1-0.3% for protein or serum concentration
- Unlikely effect sizes
  - COMT and DLPFC: meta analysis : d = 0.55, most studies N $<$ 62 subjects (Meir, 2010)

Effect size and reproducibility?

- Effect size in imaging genetics:
  - BDNF and hippocampal volume: genuine effect or winners curse? d=0.12, p=0.02, Molendijk (2012)
  - Stein et al, 2012: marker is associated with 0.58% of intracranial volume per risk allele
  - Flint 2014: Effect size of intermediate phenotype not much greater than others
  - For psychiatric diseases: mean OR is 1.15, QT: variance explained by 1 locus $<<$ 0.5%, 0.1-0.3% for protein or serum concentration
- Unlikely effect sizes
  - COMT and DLPFC: meta analysis : d = 0.55, most studies N $<$ 62 subjects (Meir, 2010)
  - HTTLPR and amygdala: Hariri 2002: p-value implies that locus explain $>$ 40% of phenotypic variance. d=1.05

Effect size and reproducibility?

- Effect size in imaging genetics:
  - BDNF and hippocampal volume: genuine effect or winners curse? d=0.12, p=0.02, Molendijk (2012)
  - Stein et al, 2012: marker is associated with 0.58% of intracranial volume per risk allele
  - Flint 2014: Effect size of intermediate phenotype not much greater than others
  - For psychiatric diseases: mean OR is 1.15, QT: variance explained by 1 locus $<<$ 0.5%, 0.1-0.3% for protein or serum concentration
- Unlikely effect sizes
  - COMT and DLPFC: meta analysis : d = 0.55, most studies N $<$ 62 subjects (Meir, 2010)
  - HTTLPR and amygdala: Hariri 2002: p-value implies that locus explain $>$ 40% of phenotypic variance. d=1.05
  - KCTD8 / cortical area: Paus 2012: 21% of phenotypic variance (250 subjects), d=1.03.

Effect size decreases with years

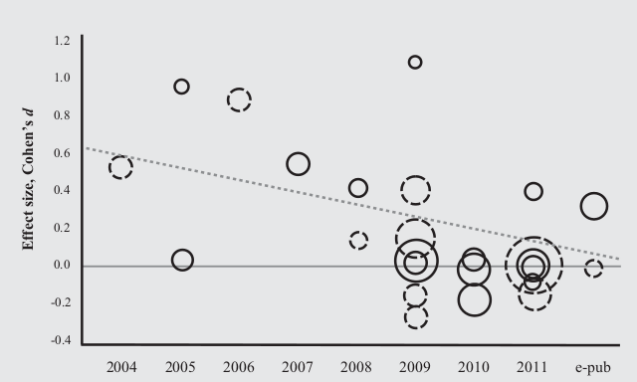


Figure 6: Molendijk, 2012, BDNF and hippocampal volume

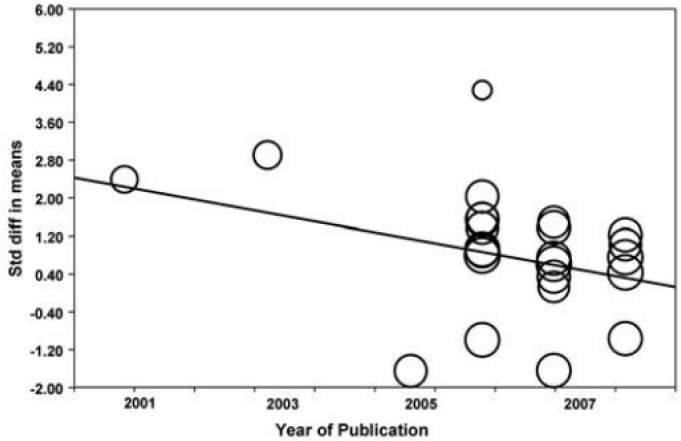


Figure 7: Mier, 2009, COMT & DLPFC

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:
  - What is your likely effect size ?

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:
  - What is your likely effect size ?
  - Power analyses with the smallest expected effect size (cost does not enter in this calculation)

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:
  - What is your likely effect size ?
  - Power analyses with the smallest expected effect size (cost does not enter in this calculation)
  - Take robust statistical tools

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:
  - What is your likely effect size ?
  - Power analyses with the smallest expected effect size (cost does not enter in this calculation)
  - Take robust statistical tools
  - Meta analysis - cf Enigma / Replication whenever possible

What are the solutions:

- Pre-register hypotheses
  - More hypotheses
  - Candidate versus GWAS: cf Flint & Mufano, 2012
- Statistics:
  - What is your likely effect size ?
  - Power analyses with the smallest expected effect size (cost does not enter in this calculation)
  - Take robust statistical tools
  - Meta analysis - cf Enigma / Replication whenever possible
  - Effect size variation estimation (bootstrapping)

Power Calculator with

- Purcell et al. "Genetic Power Calculator" Bioinformatics (2003).

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Figure 8: http://pngu.mgh.harvard.edu/~purcell/gpc/

- http://www.sph.umich.edu/csg/abecasis/cats/

CaTS-text –additive –risk 1.3 –pisample .95 –pimarkers 1. –frequency .3 –case 1067 –control 1067 –alpha 0.00000001 : yields For a one-stage study 0.314.

Recall-by-Genotype and intermediate phenotype

- Flint et al., Assessing the utility of intermediate phenotype, Trends in Neurosciences, 2014.

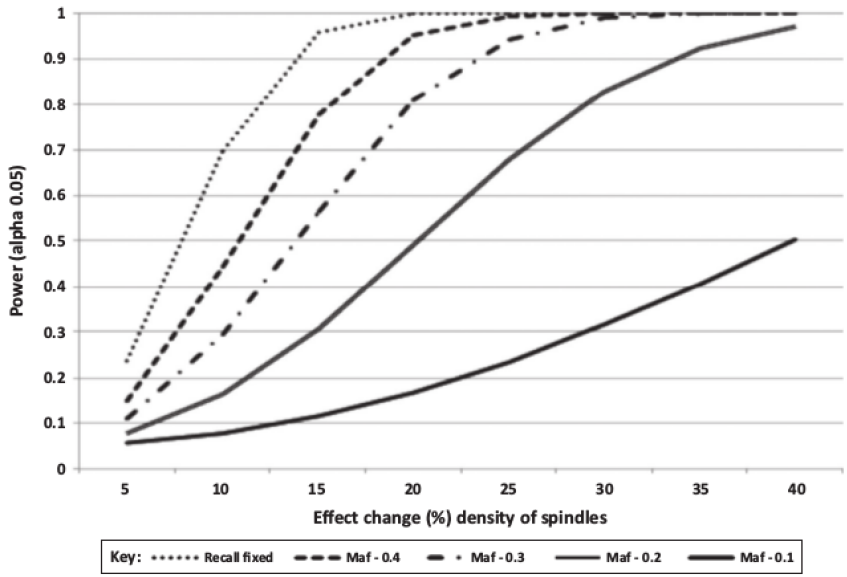


Figure 9: Recall by Genotype: Genotypic assignment vs randomisation assignment

Train the new generation

- Statistics: more in depth that what is usual.
- Computing: how to check code, version control
- A more collaborative (eg Enigma) and a more open science model (github for science)
- Work such that others in the community can reproduce **and** build upon

What are the solutions: social

- Increase awareness of editors to:
  - Accept replication studies
  - Accept preregistration
  - Increase the verifiability of analyses (code and data available)
- Share data / share intermediate results
  - Increase the capacity of the community to verify, test and re-use
  - Increase capacity to do meta/mega analyses
- Decrease publication pressure (change evaluation criteria - cf new NIH biosketch)

Conclusion : Jason's questions

1. can I publish a candidate gene study ever?
2. if I can replicate this finding with one other lab at nominal significance, is that sufficient?
3. if a SNP is genome-wide significant in a disease study, am I allowed to study its effects in my own lab without multiple comparisons correction? without replication?
4. can I study rare variants instead without worry of all this statistical correction and power?

Acknowledgement & Conclusion

- My colleagues in UCB (M. D'Esposito, M. Brett, J. Millman, F. Perez, et al.)
- My colleagues in Stanford (M. Greicius, J. Richiardi, R. Poldrack, et al.)
- My colleagues in Saclay (V. Frouin, B. Thirion)
- Jason (who reviewed all talks and had quite some work with mine :) and Tom

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.
>
> —D. Donoho