The problem of reproducibility for Imaging genetics

Jean-Baptiste Poline

Helen Wills Neuroscience Institute, UC Berkeley

Reproducibility - preliminary remarks

- Reminding ourselves : Reproducibility is the backbone of scientific activity
- Reproducibility versus replicability
- Is there a problem ?
  - Not everybody is convinced that there is a problem
  - Do we have hard evidence ?
- 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.
  - A new mechanism for independently replicating findings needed: Nature Biotech. (2012)
  - Easy to misinterpret artefacts as biologically important results: Nature (2012)
  - Too many sloppy mistakes: Nature (2012)
  - 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 Issues in Multicentre MRI Studies, Jorge Jovicich, Trento
  - Raemaekers, "Test–retest Reliability of fMRI. . . ", 2007, Thirion et al., 2007.
- 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
- Publication pressure is high
- Mistakes are done
- **Power issues**

Why do we have a problem?

Things are getting complex

- Data complexity (eg: chip idiosyncrasies, format, preprocessings, etc)
- Data need to be linked appropriately (remember the Duke scandal)
- Data size: number of variables - files you cannot check visually
- Methods: increasing number of steps and statistical complexity, external software you have to trust,

Why do we have a problem?

Publication pressure is high

- There's no way there isnt a paper out of this data set.
- You won't get your Phd if you don't publish this study
- You won't get tenure
- You won't get funding or peers recognition
- Ratio Benefice / Risk in favor of risky and quick publication
- Conclusion: the pressure is *very* high

Why do we have a problem?

Mistakes are done: unpopular topic

"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.

- Anatomy of an Error: in praise for transparency
- The Left/Right issue
- The Siemens slice ordering
- The ADHD 1000 connectomes scripts

The power issue

- Ioannidis 2005: *"Why most research findings are false"*
- Remember what is power
- What exactly 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 ?

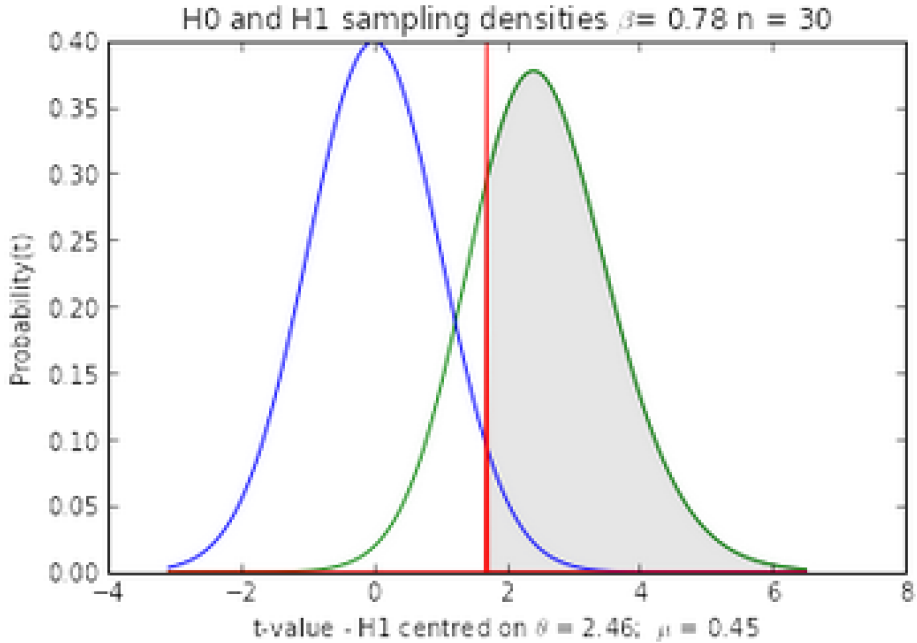

Figure: Power: $W = 1 - \beta$ Here W=77%

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

- If we have 4/5 that H0 is true, and 1/5 that H1 true, with 30% power: PPV = 60%.

$$PPV = P(H1\,True|Detection) = \frac{W\,P_1}{\alpha\,P_0 + W\,P_1}$$

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?

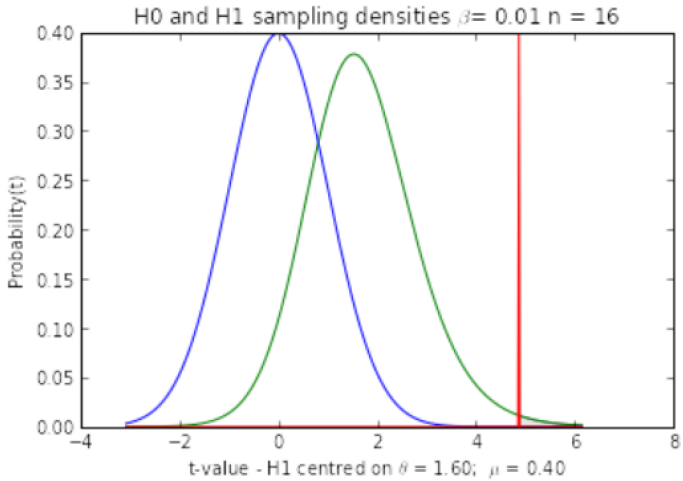

Figure: higher type I error threshold to account for MC

- 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)

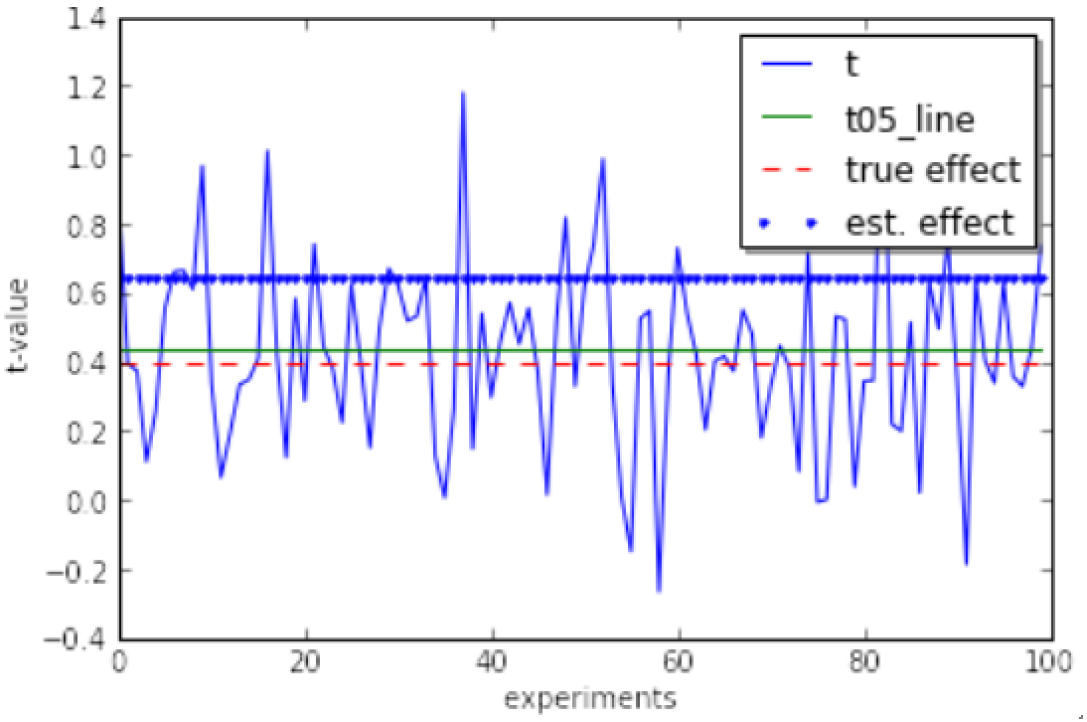

Figure: A quick simulation

The power issue

Studies of low power inflate the detected effect (1)

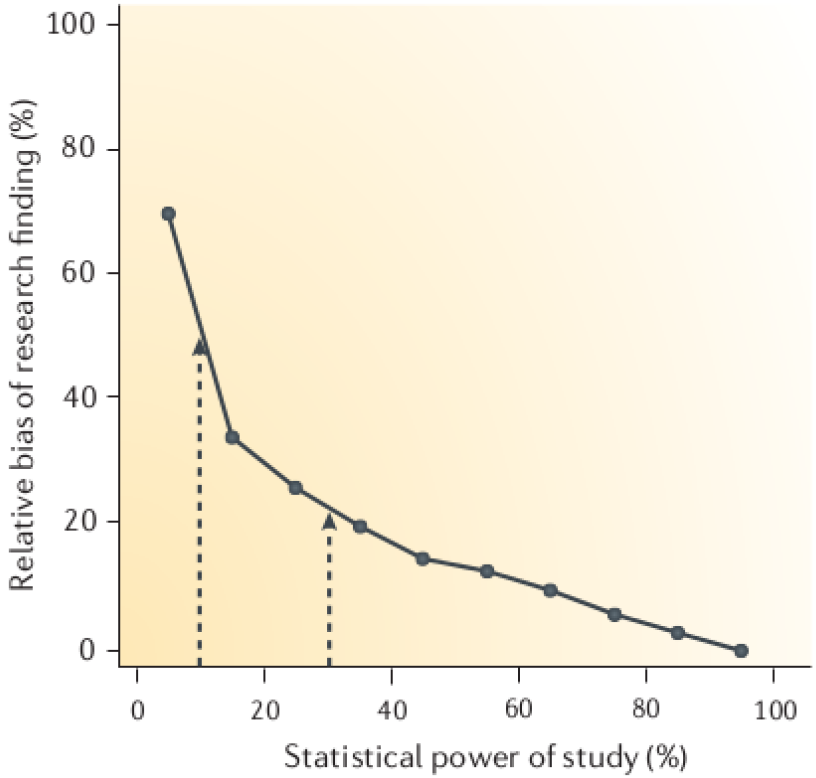

Figure: Button et al. NRN, 2013

The power issue

What is the estimated power in common meta analyses?

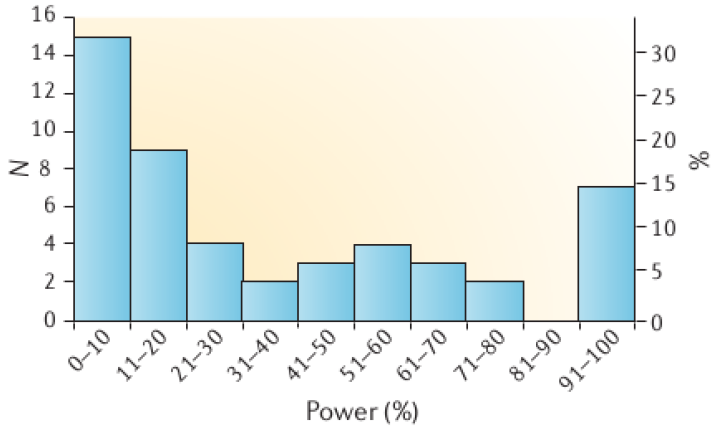

Figure: Button et al. NRN, 2013

What is specific to Imaging Genetics

- Combinaison of imaging and of genetics issues: "AND" (if independent: prob. of getting it right would multiply: .7 * .7 = .5)
- 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" (eg: noise in brain images is always interpretable)

Are imaging genetics studies reproducible?

- Effect size in imaging genetics:
  - HTTLPR and amygdala: Hariri 2002: p-value implies that locus explain about 28% of phenotypic variance.
  - KCTD8 brain growth: 2011: 21% of phenotypic variance with ˜ 250 subjects
  - BDNF and hippocampal volume : genuine effect or winners curse? d=0.12, Molendijk (2012)
  - COMT and DLPFC: meta analysis : d = 0.55, paper suggest $>$ 62 subjects Meir (2009)
  - Stein et al, 2012: rs10784502 marker is associated with 0.58% of intracranial volume per risk allele
- reproducibility / error rate
  - Silver, Montanna, Nichols (beware of low threshold forming clusters)
  - Meyer-Lindenberg et al., 2008: Not a problem ? False positives in imaging genetics. However . . .
  - Flint and Mufano: First 2002 5-HTT result is unlikely

Effect size decreases with years

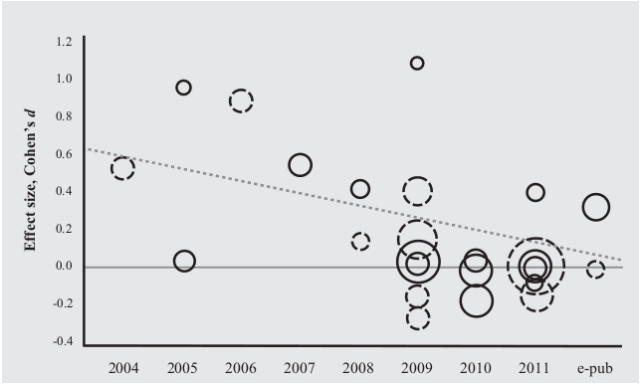

Figure: Molendijk, 2012, BDNF and hippocampal volume

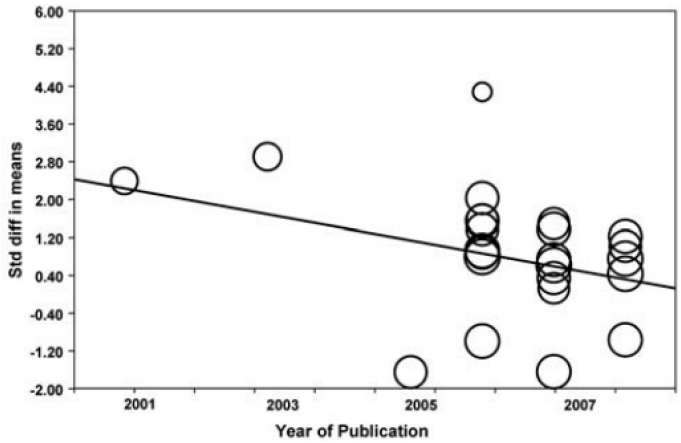

Figure: Mier, 2009, COMT & DLPFC

What are the solutions: technical

- Pre-register hypotheses
- Statistics:
  - Always try to get a sense of the power
  - Take robust statistical tools
  - Meta analysis if you can
  - Replication if you can
  - Power analyses with the smallest effect size (cost does not enter in this calculation)
  - 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: 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.

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 the next post-doc will need weeks to start to progress - not months
- 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
  - Increase capacity to do meta/mega analyses
- Change evaluation criteria - Decrease publication pressure

Acknowledgement & Conclusion

- My colleagues in UCB (M. Brett, J. Millman, F. Perez)
- 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

Figure: Donoho on publication

What are the solutions: learning

- Learn the right computing tools:
  - How can I check my code ? How can I go back to a certain state ? (learn git/mercurial, learn git Annex or others)
  - How can others check my analyses? Learn the emerging social open science frameworks
- Learn "one layer below" (A. Martelli)

[rpsychologist.com/d3/cohend]rpsychologist.com/d3/cohend