The question of reproducibility in brain imaging

Jean-Baptiste Poline

Brain Imaging Center, Helen Wills Neuroscience Institute, UC Berkeley

Outline

- Evidence for a crisis in reproducibility

Outline

- Evidence for a crisis in reproducibility
- What about brain imaging ?

Outline

- Evidence for a crisis in reproducibility
- What about brain imaging ?
- Causes

Outline

- Evidence for a crisis in reproducibility
- What about brain imaging ?
- Causes
- Impact

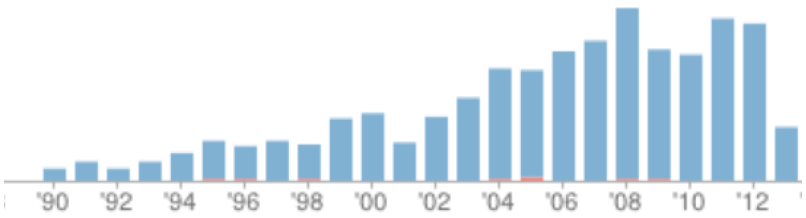
Outline

- Evidence for a crisis in reproducibility
- What about brain imaging ?
- Causes
- Impact
- What shall we do about it

Outline

- Evidence for a crisis in reproducibility
- What about brain imaging ?
- Causes
- Impact
- What shall we do about it

- Mesh terms “reproducibility of results” (100 in 2010)



Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

Genetics

Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

- Reproducible Epidemiologic Research, Peng et al, 2006.

Genetics

Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

- Reproducible Epidemiologic Research, Peng et al, 2006.
- "High FP/FN Ratio in Epidemiologic Studies", Ioannidis, 2011.

Genetics

Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

- Reproducible Epidemiologic Research, Peng et al, 2006.
- "High FP/FN Ratio in Epidemiologic Studies", Ioannidis, 2011.

Genetics

- Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: "... results are largely null."

Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

- Reproducible Epidemiologic Research, Peng et al, 2006.
- "High FP/FN Ratio in Epidemiologic Studies", Ioannidis, 2011.

Genetics

- Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: "... results are largely null."
- The failing concept of endophenotype (Iacono, Psychophysiology, 2014)

Science finding Reproducibility crisis evidence

Preclinical oncology

- Begley C.G. & Ellis L. Nature, (2012): "Only 6 out of 53 key findings in pre-clinical oncology could be fully replicated"

Epidemiology

- Reproducible Epidemiologic Research, Peng et al, 2006.
- "High FP/FN Ratio in Epidemiologic Studies", Ioannidis, 2011.

Genetics

- Ionannidis 2007: 16 SNPs hypothesized, check on 12-32k cancer/control: "... results are largely null."
- The failing concept of endophenotype (Iacono, Psychophysiology, 2014)
- Many references and warnings: eg:"Drinking from the fire hose ..." by Hunter and Kraft, 2007.

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)

Neuroscience

In general: Editorials in high profile journals

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. “P-curve”, “Evaluating Replication Results” 2014, 2013.

Neuroscience

In general: Editorials in high profile journals

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

In general: Editorials in high profile journals

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

- Nature, "Reducing our irreproducibility", 2013.

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

- Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

- Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

- Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes

In social sciences, psychology, cognitive neuroscience

- Reproducibility Project in Psychology (Open Science Framework)
- Simonshon et al. "P-curve", "Evaluating Replication Results" 2014, 2013.
- Special Issue on "Reliability and Replication in Cognitive and Affective Neuroscience Research." Barch, Deanna, and Yarkoni, Cogn Affect Behav Neurosci, 2013.

Neuroscience

- Button et al., Nature Neuroscience, 2013

In general: Editorials in high profile journals

- Nature, "Reducing our irreproducibility", 2013.
  - New mechanism for independently replicating needed
  - Easy to misinterpret artefacts as biologically important
  - Too many sloppy mistakes
- NIH plans to enhance reproducibility. Collins and Tabak, Nature, 2014.

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

- When attempted, replication is poor:

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

- When attempted, replication is poor:
  - Boekel, W., et al. (Cortex 2013) : replication study of structural brain-behavior correlations.

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

- When attempted, replication is poor:
  - Boekel, W., et al. (Cortex 2013) : replication study of structural brain-behavior correlations.
  - 5 studies, 17 findings: Bayesian analysis favored null hypothesis

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

- When attempted, replication is poor:
  - Boekel, W., et al. (Cortex 2013) : replication study of structural brain-behavior correlations.
  - 5 studies, 17 findings: Bayesian analysis favored null hypothesis
  - But: only 36 subjects, while most original studies were better powered

What about brain imaging ? Some - *but few* - facts

- Publication does not allow replication or to find methodological issues (J. Carp Cogn Affect Behav Neurosci, 2013):

"For example, while Brown and Braver (2005) claimed that activation in the anterior cingulate cortex (ACC) is sensitive to the likelihood of committing an error, Nieuwenhuis, Tanja, Mars, Botvinick, and Hajcak (2007) reported no relationship between ACC activation and error likelihood."

- When attempted, replication is poor:
  - Boekel, W., et al. (Cortex 2013) : replication study of structural brain-behavior correlations.
  - 5 studies, 17 findings: Bayesian analysis favored null hypothesis
  - But: only 36 subjects, while most original studies were better powered

- Autism example: Toro et al., Corpus callosum size example. S. Bookheimer's examples (cereb. size, FFA, FC).

Causes and Impact

Statistical

Computational

Social

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation

Example

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation
- The usual issues:

Example

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation
- The usual issues:
  - low power studies (Button et al, 2013)

Example

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation
- The usual issues:
  - low power studies (Button et al, 2013)
  - P-hacking: Simmons et al. 2011, Simmonshon et al., 2014

Example

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation
- The usual issues:
  - low power studies (Button et al, 2013)
  - P-hacking: Simmons et al. 2011, Simmonshon et al., 2014

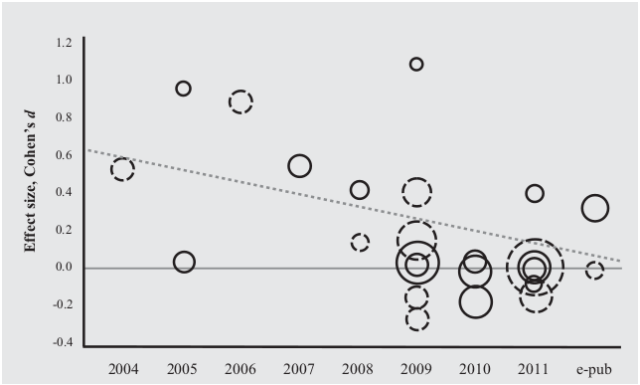
Example

- From imaging genetics (BDNF - Hippocampus volume):

Statistical causes:(1)

- Lack of understanding of statistical issues and power computation
- The usual issues:
  - low power studies (Button et al, 2013)
  - P-hacking: Simmons et al. 2011, Simmonshon et al., 2014

Example

- From imaging genetics (BDNF - Hippocampus volume):
- 

Statistical causes:(2)

- P value evil: will we eventually turn to Bayesian evidence? How ?

And if we stick to p-values:

Statistical causes:(2)

- P value evil: will we eventually turn to Bayesian evidence? How ?
- No good understanding of the necessity to report null results -> File drawer problem (Rosenthal, 1979)

And if we stick to p-values:

Statistical causes:(2)

- P value evil: will we eventually turn to Bayesian evidence? How ?
- No good understanding of the necessity to report null results -> File drawer problem (Rosenthal, 1979)
- Emmergence of complex H0/H1, A. Afraz, 2014.

And if we stick to p-values:

Statistical causes:(2)

- P value evil: will we eventually turn to Bayesian evidence? How ?
- No good understanding of the necessity to report null results -> File drawer problem (Rosenthal, 1979)
- Emmergence of complex H0/H1, A. Afraz, 2014.

And if we stick to p-values:

- Which one to pick: Revised standard for statistical evidence (PNAS Johnson 2013) : $0.05 <=> BF \in [3, 5]$

Statistical causes:(2)

- P value evil: will we eventually turn to Bayesian evidence? How ?
- No good understanding of the necessity to report null results -> File drawer problem (Rosenthal, 1979)
- Emmergence of complex H0/H1, A. Afraz, 2014.

And if we stick to p-values:

- Which one to pick: Revised standard for statistical evidence (PNAS Johnson 2013) : 0.05 $<=>$ BF$\in [3, 5]$

> **Significance**
>
> The lack of reproducibility of scientific research undermines public confidence in science and leads to the misuse of resources when researchers attempt to replicate and extend fallacious research findings. Using recent developments in Bayesian hypothesis testing, a root cause of nonreproducibility is traced to the conduct of significance tests at inappropriately high levels of significance. Modifications of common standards of evidence are proposed to reduce the rate of nonreproducibility of scientific research by a factor of 5 or greater.

-

Computational causes:

- Biologists and MDs are rarely well trained in computation - but most brain imaging findings rely heavily on computations

Computational causes:

- Biologists and MDs are rarely well trained in computation - but most brain imaging findings rely heavily on computations
- Claerbout's "" "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." ""

Computational causes:

- Biologists and MDs are rarely well trained in computation - but most brain imaging findings rely heavily on computations
- Claerbout's "" "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." ""
- Meta data capture and curation not implemented (parameters and process of data generation), no standards for meta data

Computational causes:

- Biologists and MDs are rarely well trained in computation - but most brain imaging findings rely heavily on computations
- Claerbout's "" "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." ""
- Meta data capture and curation not implemented (parameters and process of data generation), no standards for meta data
- Computational environment packaging not used (Neurodebian VM, Docker, . . .)

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies
  - favors high risk and rapid publications

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies
  - favors high risk and rapid publications
  - impeeds data and code sharing even for publicly funded research

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies
  - favors high risk and rapid publications
  - impeeds data and code sharing even for publicly funded research
- Positive finding publication bias and the file drawer problem

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies
  - favors high risk and rapid publications
  - impeeds data and code sharing even for publicly funded research
- Positive finding publication bias and the file drawer problem
  - how this can delay scientific revolution (A. Afraz, 'We could all be astronomers')

Social/systemic causes

- Publication based reward system (career, grants, fame, etc) + hyper competitive environment is in general not working towards good science:
  - favors speed over careful, **re-usable**, reproduced studies
  - favors high risk and rapid publications
  - impeeds data and code sharing even for publicly funded research
- Positive finding publication bias and the file drawer problem
  - how this can delay scientific revolution (A. Afraz, 'We could all be astronomers')
  - is science always self-correcting ?

Impact

- Large amount of resource wasted (talent, money, time)

Impact

- Large amount of resource wasted (talent, money, time)
- Discredit from the public and governments

Impact

- Large amount of resource wasted (talent, money, time)
- Discredit from the public and governments
- Slows down scientific and medical progress

Impact

- Large amount of resource wasted (talent, money, time)
- Discredit from the public and governments
- Slows down scientific and medical progress
- Impact on the type of work that can be started (counter example: UK biobank, Bavarian cohorts).

Impact

- Large amount of resource wasted (talent, money, time)
- Discredit from the public and governments
- Slows down scientific and medical progress
- Impact on the type of work that can be started (counter example: UK biobank, Bavarian cohorts).
- The system may select the most "productive" scientists - not necessarily the best

Conclusion: What shall we do about it

- Adopt more stringent and better statistical and computational standards

Conclusion: What shall we do about it

- Adopt more stringent and better statistical and computational standards
- Adopt genetic research standards for replication

Conclusion: What shall we do about it

- Adopt more stringent and better statistical and computational standards
- Adopt genetic research standards for replication
- Adopt clinical trial standards and pre-registration

Conclusion: What shall we do about it

- Adopt more stringent and better statistical and computational standards
- Adopt genetic research standards for replication
- Adopt clinical trial standards and pre-registration
- Augment the awareness of these issues, adopt data and code sharing as the standard in our field

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists
  - online forum for open discussions,

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists
  - online forum for open discussions,
  - change in funding and bio, anonymize peer review, etc.

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists
  - online forum for open discussions,
  - change in funding and bio, anonymize peer review, etc.
- Work with journals and editors to accept well powered null findings

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists
  - online forum for open discussions,
  - change in funding and bio, anonymize peer review, etc.
- Work with journals and editors to accept well powered null findings
- OSF, many lab and community based projects, METRICS (Meta-Research) institutes

Conclusion: What shall we do about it

- Train the new generation of scientist in computation, statistics
- NIH answers:
  - Data Discovery Index, checklists
  - online forum for open discussions,
  - change in funding and bio, anonymize peer review, etc.
- Work with journals and editors to accept well powered null findings
- OSF, many lab and community based projects, METRICS (Meta-Research) institutes
- Reward people who produce re-usable science

Acknowledgement

- At Berkeley: M. Brett, J. Millman, F. Perez; Simpace interest group: D. Sheltraw, C. Gallen, A. Tambini, K. K. Hwang; B. Inglis, M. D'Esposito.
- At INCF: Mathew Abrams, Linda Layon, Roman Valls
- Nidash: David Kennedy, Satra Ghosh, Chris Gorgowleski, Nolan Nichols, Dave Keator, Camille Maumet, Guillaume Flandin, Tom Nichols, Russ Poldrack, etc
- At Pasteur, Neurospin, MNI.