OPEN DATA

Sustainability and transparency in computational cognitive neuroscience
OPEN DATA

Quality improvement
• Reproducibility, methodological rigor, error discovery, ...

Cost reduction
• Redundancy decrease, resource re-allocation, ...

Interdisciplinarity
• Mathematics, statistics, computer science, ...

Progress acceleration
• Meta-analyses, multi-site generalization, power, .....

Zeitgeist
• Funding agency demands, ...

⇒ Cognitive neuroimaging has to adopt an open science standard!

Poline et al. 2012; Brakewood & Poldrack, 2013
Outline

Community efforts: COBIDAS and BIDS

Ethical and legal challenges

Scientific common sense

MUST

SHOULD

DISCUSS
Outline

Community efforts: COBIDAS and BIDS

Ethical and legal challenges

Scientific common sense
COBIDAS background

Preemptive reaction of OHBM to perceived „reproducibility crisis“

OHBM Council Statement on Neuroimaging Research and Data Integrity (2014)

Committee on Best Practices in Data Analysis and Sharing (COBIDAS)

• Identify best practices of data analysis and data sharing in brain mapping

• Preparation of a white paper organizing and describing these practices

• Seeking input from the OHBM community

• Publishing the recommendations
COBIDAS background

Committee on Best Practices in Data Analysis and Sharing (COBIDAS)

• „Proposed“ version of the report posted in October 2015

• Comments collected via the COBIDAS blog (https://cobidas.wordpress.com/)

• Ratification in May 2016 with 96% positive ballots

• Voting turnout 153 yes, 6 no (≈3,000 attendees of OHBM conference/year)
COBIDAS background

Committee on Best Practices in Data Analysis and Sharing (COBIDAS)

• Increase transparency by comprehensive sharing of data, methods, and results

• Recommendations and checklists to increase reproducibility

• Tool, not law!

• Poldrack et al. (2008) Guidelines for reporting an fMRI study *Neuroimage*
COBIDAS background

Defining reproducibility

**ISO* Repeatability**
Precision of measurements under conditions where independent test/measurements results are obtained with the same method on identical test/measurement items in the same test or measuring facility by the same operator using the same equipment within short intervals of time.

**ISO Reproducibility**
Precision of measurements under conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in different test or measurement facilities with different operators using different equipment.

**Computational Reproducibility**
Independent researchers use the exact same data and code to arrive at the original result (Peng (2011) Reproducible research in computational science, *Science*).

**Replicability**
Independent researchers use independent data and possibly distinct methods to arrive at the same original conclusion.

*International Standards Organization*
COBIDAS content

1) Experimental design reporting
2) Acquisition reporting
3) Preprocessing reporting
4) Statistical modelling & inference
5) Results reporting
6) Data sharing
7) Reproducibility
COBIDAS content

1) Experimental design reporting

**Box 2.1. fMRI Terminology**

**Session.** The experimental session encompasses the time that the subject enters the scanner until they leave the scanner. This will usually include multiple scanning runs with different pulse sequences, including structural, diffusion imaging, functional MRI, spectroscopy, etc.

**Run.** A run is a period of temporally continuous data acquisition using a single pulse sequence.

**Volume.** A volume (or alternatively “frame”) is single 3-dimensional image acquired as part of a run.

**Condition.** A condition is a set of task features that are created to engage a particular mental state.

**Trial.** A trial (or alternatively “event”) is a temporally isolated period during which a particular condition is presented, or a specific behavior is observed.

**Event.** The term “trial” and “event” are often interchangeable. However, in certain situations of ‘compound-trials,’ a trial will consist of multiple subunits; for example, a working memory task may contain subunits of encoding, delay, and retrieval. In these cases the subunits are labeled as “events” and the “trial” is defined as the overarching task.

**Block.** A block (or alternatively “epoch”) is a temporally contiguous period when a subject is presented with a particular condition.
COBIDAS content

2) Acquisition reporting
3) Preprocessing reporting
4) Statistical modelling & inference
5) Results reporting

Know exactly what your tool is doing!

Transparency = Methodological knowledge
6) Data sharing

... does not mean „available upon request“

... should adhere to the FAIR principles and use a standardized data format

... should be an essential project component (data management plan)

... may use appropriate licenses, such as Creative Commons

... should include data analysis pipeline scripts

... should happen at the time of study publication

... should respect the appropriate ethical and legal constraints
7) Reproducibility

... stresses „computational reproducibility“

... requires the analysis scripts to be documented, archived, and citable

... may involve creation and release of virtual machines (VM)

... should us Digital Object Identifiers for permanent citability
COBIDAS summary

- Best practices in data analysis and sharing in neuroimaging using MRI
  - [www.biorxiv.org/content/biorxiv/early/2016/07/10/054262.full.pdf](http://www.biorxiv.org/content/biorxiv/early/2016/07/10/054262.full.pdf)

- A good starting point for writing neuroimaging methods sections

- Fairly silent on practices of data sharing

- Relatively low community engagement
Outline

Community efforts: COBIDAS and BIDS

Ethical and legal constraints: GDPR

Scientific common sense: Personal views
BIDS

BIDS = Brain Imaging Data Structure
A standard for organizing data of human neuroimaging experiments


Gorgolewski, K, ... , Poldrack, R (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments *Scientific Data* Volume 3, Article number: 160044
BIDS

Core principles

- Foster wide adoption by simplicity and familiarity (openfMRI)
- Some metadata is better than no metadata
- Independence of external software or complicated file formats
- Capture 80% of experiments, but allow extension for the rest
BIDS

Implementation

- Some metadata encoded in folder structure and replicated in filenames
- Tab separated files for tabular data
- Compressed NIFTI files for neuroimaging data
- JSON files for dictionary-type metadata
- Text file formats for behavioral and physiological data
- Few musts, many can dos
Folder organization

sub-control001/
  anat/
    sub-control001_T1w.nii.gz
    sub-control001_T1w.json
  func/
    sub-control001_nback_bold_01.nii.gz
    sub-control001_nback_bold_01.json
    sub-control001_nback_bold_02.nii.gz
    sub-control001_nback_bold_02.json

← Participant data folder
← Anatomical data folder
  ← Data
  ← Metadata
← BOLD data folder
  ← Data
  ← Metadata
BIDS

Folder organization
<table>
<thead>
<tr>
<th>Onset</th>
<th>Duration</th>
<th>Trial_Type</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>0.7</td>
<td>easy</td>
<td>1.24</td>
</tr>
<tr>
<td>5.6</td>
<td>0.7</td>
<td>hard</td>
<td>2.10</td>
</tr>
<tr>
<td>9.2</td>
<td>0.7</td>
<td>hard</td>
<td>1.92</td>
</tr>
<tr>
<td>11.1</td>
<td>0.7</td>
<td>easy</td>
<td>0.98</td>
</tr>
<tr>
<td>15.0</td>
<td>0.7</td>
<td>easy</td>
<td>1.34</td>
</tr>
<tr>
<td>18.3</td>
<td>0.7</td>
<td>hard</td>
<td>1.97</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
BIDS

Metadata JSON file (key-value data/structure format)

{

   "Repetition Time" : 2,
   "EchoTime" : 0.03,
   "FlipAngle" : 79,
   "Slice Timing" : [0.0, 0.2, ..., 1.8, 2.0],
   "MultibandAcceleration" : 4,

   ...
   ...

}
Software currently supporting BIDS:

- **BIDS Apps** (a growing set of portable containerized data processing pipelines that understand BIDS datasets)
- **Converters**
  - AFNI BIDS-tools
  - BIDS2SATab
  - BIDSTo3col
  - BIDS2NDA
  - bidskit
  - dac2bids
  - Dem2Bids
  - DCM2NIx
  - DiCM2NII
  - HeuDiConv
  - OpenfMRIBIDS
  - Reproln (HeuDiConv-based turnkey solution)
  - bids2xar (for XNAT import)
  - XNAT2BIDS
  - Horos (Osirix) export plugin
  - BIDS2NIDM
  - MNE-BIDS (MEG/EEG/iEEG)
- **Institution specific data management/conversion tools**
  - BIDS Tools from the Donders Institute
  - Autobids from the Centre for Functional and Metabolic Mapping (CFMM) at Western's Robarts Research Institute
- **Other Tools**
  - Automatic Analysis (fMRI processing toolbox)
  - Brainstorm (MEG/EEG analysis package)
  - C-PAC (Configurable Pipeline for the Analysing Connectomes)
  - FMRIprep (preprocessing workflow)
  - OpenNeuro (repository)
  - PyBIDS (Python module to harmonize access and manipulation)
- **Quality Assessment**
  - MRIQC
  - QAP

A description of how to build containerized apps supporting BIDS inputs can be found in the paper published in PLOS Computational Biology.
Summary

JUST DO IT.
Outline

Community efforts: COBIDAS and BIDS

Ethical and legal challenges

Scientific common sense
Ethical and legal challenges
Ethical and legal challenges

Data privacy protection

Data sharing in basic research

Our task!
Ethical and legal challenges

European General Data Protection Regulation

What is it?

• Regulation 2016/679 by EU parliament, council, commission
• Legally binding and applicable from May 25th, 2018
• EU reaction to global surveillance disclosures

What are its aims?

• Strengthen and unify data protection for individuals within EU
• Empower citizens in control of their digital data
Ethical and legal challenges

**European General Data Protection Regulation**

What is its relation to neuroimaging data sharing?

- GDPR applies to EU data collectors, processors, or subjects

- Personal data := Any information relating to an individual
  - Name, address, photo, medical information, posts, ...

- Sanctions and fines

- Designed independently of open science principles
Ethical and legal challenges

European General Data Protection Regulation

What are its main principles?

• Processing of personal data is prohibited, unless consent is provided
• Data collection only for specified, explicit, and legitimate purposes
• Data processing has to ensure appropriate security
• The data controller is responsible for GDPR compliance
• Science has a degree of privilege for processing personal data

[GDPR Articles 4, 5, 9, 89]
Ethical and legal challenges

Elena Pavlenko & Dirk Ostwald

Cognitive Neuroimaging Data Sharing: A European Perspective

Participant consent
Data storage
Data sharing

• What does the GDPR specify?
• What does the GDPR not specify?

Cognitive Neuroimaging Data Lifecycle
Ethical and legal challenges

• The GDPR does not apply to anonymous data

• Can functional neuroimaging data be anonymized?

• “... [Anonymous data is] information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. [...] To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.”

[GDPR Article 4, Recital 26]
Ethical and legal challenges

Anonymity and anatomical data

Anonymity/Identifiability and functional (BOLD) data (e.g. Finn et al. (2015))

Can functional neuroimaging data be anonymized?
Ethical and legal challenges

Some GDPR legal uncertainties for cognitive neuroimaging

• Can functional neuroimaging data be anonymized?

• How broad can the purpose of data processing be specified?

• How specific do possible data users be named?

• How can participant rights be maintained upon data sharing?

• Can participant consents partially overrule participant right?
Open data modes

- Public Sharing
- Restricted Sharing
- Dynamic Sharing
Open data modes

Unrestricted access for anyone with an internet connection

**Advantages**
- Simplicity
- Objective data usage rights
- Small cost and effort
- Data processing for all purposes

**Disadvantages**
- No form of data usage control
- GDPR-compatible consent difficult
- Risk of data abuse
- Data processing for all purposes
Open data modes

Restricted Sharing

Researcher-regulated password protected data access + data use agreement

Advantages

• Researcher in control
• Low risk of data abuse
• Subjective data usage rights
• Data use agreements

Disadvantages

• Researcher effort
• Researcher responsibility
• Subjective data usage rights
• Ressource intensive
Open data modes

Researchers-regulated password protected data access + data use agreement
Open data modes

Participant-regulated password protected data access + data use agreement

Advantages

• Empowerment
• Subjective data usage rights
• Researchers not in charge
• Participant engagement

Disadvantages

• Digital divide
• Subjective data usage right
• Effort for participants
• Ressource intensive

⇒ 21st century value of and responsibility for one’s own digital data
Open data modes

- Public Sharing: OpenfMRI
- Restricted Sharing: OSF
- Dynamic Sharing: MIDATA

Max Planck
Outline

Community efforts: COBIDAS and BIDS

Ethical and legal constraints: GDPR

Scientific common sense
a.k.a. my personal views

Sustainability in (computational) cognitive neuroscience

- Day-to-day practice
- Strategic choices
- Scientific philosophy
Scientific common sense

- **Day-to-day practice**
  - Use standard data organization
  - Do not use toolboxes for central analyses
  - Write clean, understandable code
  - Set up result-independent win-win situations
  - Sign your reviews
  - Don’t believe in papers, because they are peer-reviewed
  - Acquire the language to talk about computational models
Scientific common sense

- **Strategic decisions**
  - Work with supervisors who work transparently
  - Start university teaching
  - Emphasize *Science* over *Academia*
  - Shape the open research movement
Scientific common sense

Scientific philosophy

- Science is communication, not the pursuit of absolute truth
- There is no “correct” statistical model or analysis
- Embrace uncertainty and modesty
- Don’t buy into the competitiveness noise
Scientific common sense

Scientific philosophy

Reality → Science → Communication

Data

Model

Result

„Statistics“
„Probability modelling“
„Multivariate models“
„Dynamical models“
„Quantitative theories“
...

\[ \dot{x}(t) = ax(t) \]
\[ \hat{a} = 2.1 \]
Summary

Community efforts: COBIDAS and BIDS

Ethical and legal challenges

Scientific common sense
Interdisciplinary Perspectives on Open Science and Open Scholarship

**Discussion and Networking Event**

with contributions by

- Arno Mährle (Institute of Limnology)
- Claudia Müller (Institute of Computer Science)
- Dirk Ostwald (Institute of Education and Psychology)
- Cornelia Rüfener (Department of History and Cultural Studies)
- Christian Winterreuther (University Library)
- Agnieszka Wereszczak (Center for Digital Systems)

June 6th 2019, 6 pm, Room L115, Seminarzentrum

**Berlin Institute for Health Research**

BIH QUEST CENTER FOR TRANSFORMING BIOMEDICAL RESEARCH

**Quest – Quality | Ethics | Open Science | Translation**

**Berlin | Oxford Summer School on Open, Transparent, and Reproducible Research in the Life Sciences**

September 2019