major problems for image metadata variation & infrastructures

Jean-Baptiste Poline
UC Berkeley
CEA Neurospin

Outline

- Problems start before data acquisition
- During data acquisition : the capture of meta data
- The meta data exchange issues
- Infrastructure for sharing
- Infrastructure for discovery
- Statistical issues
- Sustainability / Reproducibility issues

Problems start before data acquisition

- **Ethical aspects**
  - Standardize informed consent / data stewardship vs ownership / Legal variations across countries
  - Agree on the data use policy – check international standards for data sharing and authorship
  - Ethical rationale for open data within privacy protections
  - Standard Guid from NIH / Centre TBI
- **Study design** : power analyses / sampling questions
- **Acquisition Standardisation issues**
  - The problem of the minimal requirements across scanners
  - Standardize versus capture and model variability
    - Sensitivity versus generalizability

Acquisition of meta data

- Imaging meta data: rely on dicom extraction
  - Dicom terms in Neurolex for definition / Nifti issues
- What tools do we have to capture meta data during/around acquisition
  - Not electronic tools : risk of error
  - Some electronic lab notebooks (ELNs)
  - Many online instruments
    - Few taxonomy and ontologies around these instruments
- More complex data/MD : stimuli & timing,
  - Physiological data
  - Some work in NIMS

The meta data exchange issues

- Controled vocabulary are often project dependant
- Definition rarely provided - Already existing lexicon / ontologies re-used - uri not dereferencing
- Neurolex

- A much lesser problem : Format of exchange may vary (ascii, xml, json, excel, …)

DB imaging Infrastructures

- XNAT, LORIS, COINS, NIMS, HID, LDA, CubW, FIPS, NIDB, XXX, …
  - Simplicity, flexibility, maintenance, support,
- Often linked to pipelining systems
- Often centralized system
  - Local instances installed – future in web technologies?
- Capacity to federate / include heterogeneous data
  - behavioural, clinical, genetics, but also implementation of project management and sharing policies

Issues

- **Data versioning** : **When and what to release ?**
  - Raw data / Preprocessed / processed data / QC
- **Duplication of data** on several local : data unique identifier : The NIF problem
- **Data discovery** across infrastructures : queryable meta data – common API
- **Infrastructure maintenance after funding**
- **Local file system sync**– DB upload / download
  - Eg PyXnat / discover when data have changed
- **Move computation** to the data – pipelines provenance
- Share pipelines and statistical results

Infrastructure for discovery

- NIF
- Nitrc
- Google ...
- Future : **W3C Prov model**

QC/QA

- Distinguish between
  - Right format / QC that need only this one piece of data – automatic vs manual
  - Data QC that requires distributions
- QC depends on study use
- New QC measures every day
- Multivariate aspects
- No standardization across projects : but large projects examples (eg fBirn) – **one click tool**

Statistical issues

- Variability across sites : assess / correct
  - Population, SOP, scanner hardware, sequence
  - Correction and models : mixed effects / meta analyses / stratification issues
- **INCF NIDM as a standard for results meta data**
- Larger statistical issues
  - Accounting for previous results on the same data
    - Corrections for Increased alpha risk -
    - Bayesian analyses / predictive models
  - Recording previous results / FDR on research findings ?
  - Preregistration of hypothesis

Reproducibility Sustainability

- Reproducing results from very large datasets ?
  - Power can still be an issue (eg img/genetics)
- What economical model for sustainability of the infrastructurs
  - The « more grants » model
  - Acquisition percentage from funding agencies
  - Other organisations

Thanks

- INCF Neuroimaging data sharing group (NiDash)
- Neurospin colleagues
- Berkeley colleagues
- Imagen colleagues