Ethics of FAIR in neuroimaging

Gustav Nilsonne

2021-04-20

@GustavNilsonne

Outline

- FAIR vs open data
- Weighing benefits against risks and harms
- Assessing reidentification risk and sensitivity of data
- Synthetic data as a possible solution
- Recommendations for information to participants
- Some discussion points

FAIR vs open data

- FAIR: As open as possible, as closed as necessary
- Data sharing enables
  - → Cumulative science
  - → New discoveries
  - → Critical appraisal

Weighing benefits against risks and harms

Benefits
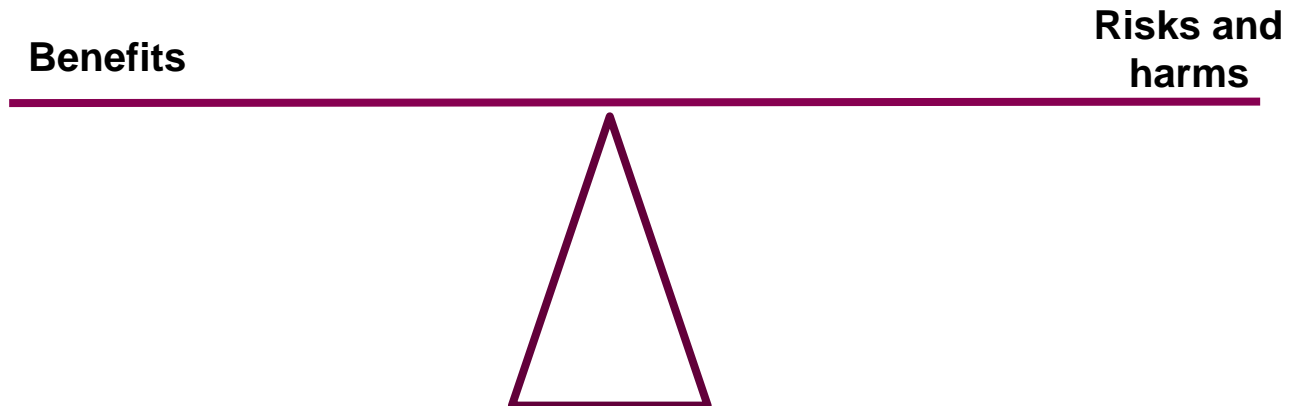
Risks and harms

Weighing benefits against risks and harms

- We have to protect our participants and their privacy
- We also have to make sure that the greatest knowledge gain can be obtained from their participation (e.g. Brakewood & Poldrack)

Assessing risk of reidentification

- Risk depends on **likelihood of reidentification** and **sensitivity of data**
- What kinds of data are included in the dataset?
  - → Health information, sexual orientation, genetic information etc
- Threat models
  - → Self-identification by participant
  - → Targeted reidentification
  - → Mass reidentification
- Data uniqueness: can a participant be singled out?

| Target population | ←→ | Reference population |
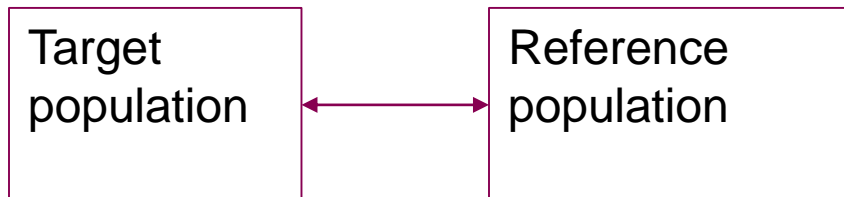
Risk assessment: examples

- T1/T2 images (defaced)
  - → Can be matched to other structural brain images
  - → Possible reference datasets include clinical data and other research datasets

- Raw functional images
  - → Can be matched by anatomy to other brain images
  - → Possible reference datasets include other research datasets

- Normalised functional images
  - → Can be matched by activation patterns to other images (success rate is questionable)
  - → Possible reference datasets include other research datasets
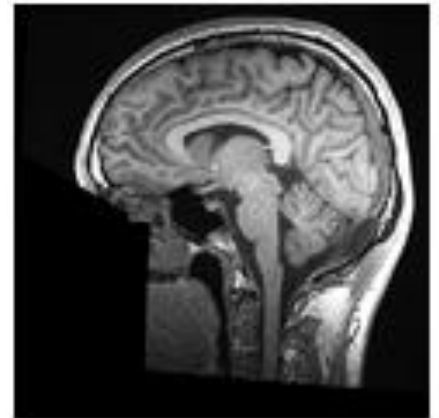
Synthetic data: a possible solution

- Synthetic/simulated data can be freely shared if the data generating procedure ensures personal information is not revealed

- May be very useful e.g. for exploration and model building

- Inference often requires real data, to ensure conclusions are not based on properties introduced by the simulation
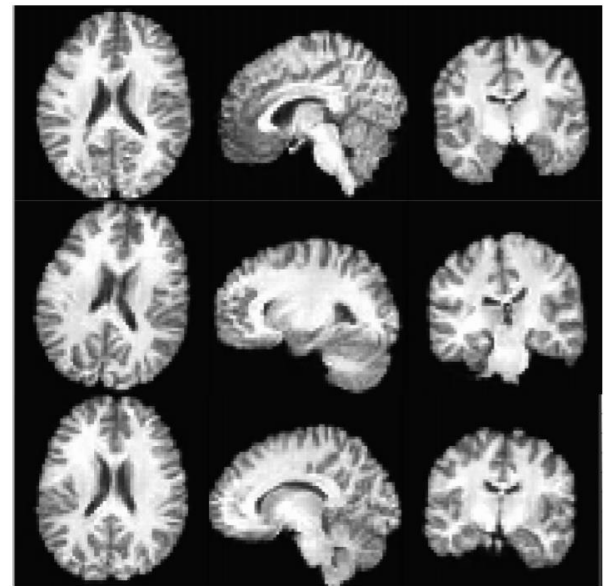
Brain images synthesised by a generative adversarial Network (GAN); Eklund 2019

Recommendations for information to participants

Consider phrases like:

- "Data will be published openly in the XX archive [link]"
- "We will remove all data we think could be used to identify you in the published data set, for example name and date of participation"

Avoid phrases like:

- "No-one outside the research group will have access to your data"
- "Results will be published only as statistical averages"
- "Your data will be stored for ten years"

Open Brain Consent

- Example language in many languages available at https://open-brain-consent.readthedocs.io/en/master/#

Some discussion points

- Incentives for sharing: how can we make FAIR data the norm?
- Regulatory obstacles
    - → Increased jurisdictional siloing
    - → Lack of consensus in the EU on what constitutes personal data
- Federated analyses: a possible solution?
    - → Difficult and expensive to operate