







Protection of Privacy – perspective from the UK

Clare Mackay University of Oxford



NAS- Forum on Neuroscience and Nervous System Disorders Neuroscience Data in the Cloud– a Workshop September 24, 2019

Distinct use cases: one size does not fit all

Single cohort/study



- Multiple data types
- Single IRB & consent
- Single or multiple institutions
- Single PI/ data access procedure

Individual facility



- Restricted data types
- Multiple study types, IRBs, consents
- Multiple PIs, multiple funders
- Single institution

(Sub)Field





- Restricted data types
- Multiple study types, IRBs, consents
- Multiple Pis, multiple funders
- Multiple institution



Single study – UK Biobank example

- Population cohort of 500,000 individuals, open by design
- 100K currently undergoing imaging sub-study (45K already scanned)
- Data includes demographics, extensive health questionnaire, blood sample, genetics, touch-screen assessments of cognition
- Data access:
 - Individuals apply (and pay admin fee) for access 2 stage process
 - Data access committee
 - Distinction between sensitive and non-sensitive data
 - No disclosure
 - Bespoke arrangements with institutions for local copies of the data

Single facility – WIN example



- Imaging facility, approx. 35 PIs
- Studies of all shapes and sizes (basic, clinical, commercial)
- Data includes demographics, MRI, MEG, PET, electophys, behavioural data
- Data access:
 - Developing centralized infrastructure to facilitate data sharing
 - Responsibility for privacy remains with the study PI



Field platform



Layer

Data

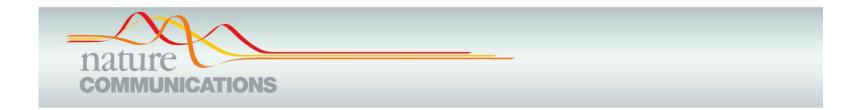
extract

Centrali



- 35 cohorts, >3M participants
- Platform infrastructure to facilitate data sharing, data aggregation and cross-cohort analyses
- 'Single point of entry' data access request system, but responsibility lies with the cohort PI

Layer 1: Data Inges	stion / Remote access		Layer 4	l: Data Discov	erability	
Data tract Data extract EHR Di	Data extract Data extract gital Research		ery tools	Queries	APIs	ļ
linkage Phen	otyping readiness registers	-7	ayer 5. Data/		ccess Brokering	
Layer 2: Data Curation		Fail-fast a criter		ndard legal agreements	Recontact for clinical studies	Ļ
Standardisation + Harmonisation APIs			Layer 6: Data Analysis			
Curated Data Assigned DOIs	Native Data Assigned DOIs	Analysis	られ Personal workspaces	Consortium workspaces	APIs Containers	ļ
Layer 3: Interoperability			Layer 7: Knowledge Management			
Interface with other Platforms	By-design security for federated analyses	Catalogu system	-	wledge graph: lligent learning	Knowledge Preservation: Community Archive	ļ



ARTICLE https://doi.org/10.1038/s41467-019-10933-3 OPEN Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher^{1,2,3}, Julien M. Hendrickx¹ & Yves-Alexandre de Montjoye^{2,3}

heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-andforget model.

Perspective of participants

• Privacy, security, yes but...

• 'Informed consent'



• Reindentification: Rights/opportunities of cohort participants

Hard (but vitally important) sells

- Risk-benefit
- Industry partnerships are good for us all!
- Fear of the known/unknown unknowns
- Power to the participants

