Global Genomic Medicine Collaborative

International Cohorts Summit

Breakout Group 2

IT considerations for enabling coordination, communication, centralization (include federated databases)

Define the Cohorts

Will need to ascertain:

- What data are collected and available for each participating cohort and in what format
- What are sources of those data (e.g. electronic health record, laboratory, radiology, genomics, proteomics, metabolomics, others)
- How data collected in each cohort are stored

Define the Cohorts

Will need to ascertain:

- What data are collected and available for each participating cohort and in what format
- What are sources of those data (e.g. electronic health record, laboratory, radiology, genomics, proteomics, metabolomics, others)
- **How data collected in each cohort are stored:**
  - **How that varies by data type (genomic, electronic health record data, metadata, other)**
  - **Benefits and drawbacks of approaches being used**
  - **Standards being used to store the data or metadata**

Federation vs Centralisation of (data) portals

International Cohorts Summit

Global Genomic Medicine Collaborative

- Which model is suitable for large scale cohorts?

- Centralisation
  - Reduce duplication of effort
  - Few single point for data discovery+access
  - More sustainable funding? Easier to get funding for few much much larger portals?

- Reasons for federation
  - Jurisdictional restrictions on export, esp. healthcare data
  - Data safe haven
  - Enhanced security requirements
  - Sheer size of genomic datasets, not practical to copy
  - Challenge: Sustainability of many smaller (data) portals

Federated Analysis Model

International Cohorts Summit

Global Genomic Medicine Collaborative

- Pharma already works in this model
  - Request access, login to another infrastructure, analyse data, export allowed results
  - Different from how cohort research is currently carried out

- Standard interfaces required
  - UK, DataShield
  - UK100K project - GA4GH driver project/Cloud workstream
  - **Want to avoid having to write specific pipeline to run analysis at every site**

Interoperable analysis

International Cohorts Summit

Global Genomic Medicine Collaborative

- What does it look like?
  - Standard set of analysis modules to choose from at each cohort
- Goal of GA4GH cloud workstream
- Heterogeneity of tools
  - Will this always be the case for genomics?
  - Simply acknowledge the situation
  - Standardisation of analysis modules may help to address this

PanCancer - one big variable was the sample collection artefacts rather than the analysis differences

Authorisation and Access

International Cohorts Summit

Global Genomic Medicine Collaborative

- Goal is to accelerate the authorisation process
  - Scenario: apply for authorisation for dataset 1 at institute X, go through full authorisation process, then requests for other datasets at institute X can be expedited

- Researcher Library card
  - Researcher gains access to other resources based on their already having access to other datasets
  - Researcher ID/passport

- Bona-fide researcher
  - What is it? How is it determined?
  - Can we have a universal definition?

Operationalising Consent

Global Genomic Medicine Collaborative

Federation model and the original consent

Funding required to operationalise, harmonise, and standardise data and metadata

- Is this achievable? Define baseline set of goals
- Cross harmonisation of multiple cohorts

Duty of care for the operator/provider

- Ensure all access is compliant with the participant consent
- Harmonisation challenge
- Wide variety of models, this is a reality

Legacy data sets with legacy IT systems

Longitudinal cohort data has been collected for 20-40 years

- Heterogeneity of IT systems, and data collection protocols

How to integrate with modern federated systems?

Integration examples

- cdisc format - pharma data has to be converted into the format
- At a similar point for cohort studies

Data/metadata Harmonisation

- Metadata - ontology mapping tools exist, cohort level metadata harmonisation,
- How you ask the questions to participants, beyond yes/no questions

Discoverability

International Cohorts Summit

Global Genomic Medicine Collaborative

Searchable metadata:

- Cohort level (eg sample description, requirements for data use, sample processing)
- Broad map of content (ontology)
- Detail about individual variables (eg sample size, equivalence across waves)

GA4GH Discovery Workstream

Beacon Network

- Discovery by genotype - Beacon network
- Discovery by phenotype - Matchmaker exchange
- Discovery by data use - Data Use Ontology (see later slide)

Matchmaker Exchange

Genomic discovery through the exchange of phenotypic & genotypic profiles

Ontologies for harmonisation

International Cohorts Summit

Global Genomic Medicine Collaborative

Data and metadata harmonisation is one of the most significant challenges for long term cohorts

Ontologies

- Formal naming and definition of the types, properties, and interrelationships of the entities

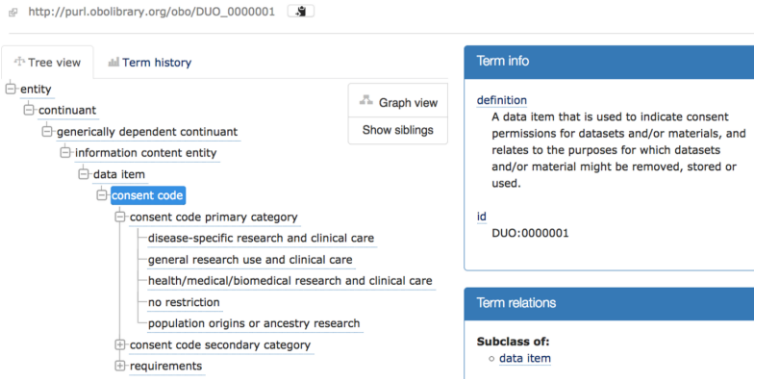
Examples

- Data Use Ontology (DUO)
  - Semantically tag datasets with restriction about their usage

Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke, Anthony A. Philippakis, Jordi Rambla De Argila, Dina N. Paltoo, Erin S. Luetkemeier, Bartha M. Knoppers, Anthony J. Brookes, J. Dylan Spalding, Mark Thompson, Marco Roos, Kym M. Boycott, Michael Bru Matthew Hurles, [···], Stephen T. Sherry [view all]

Published: January 21, 2016 • https://doi.org/10.1371/journal.pgen.1005772



Hypothesis free research

International Cohorts Summit

Global Genomic Medicine Collaborative

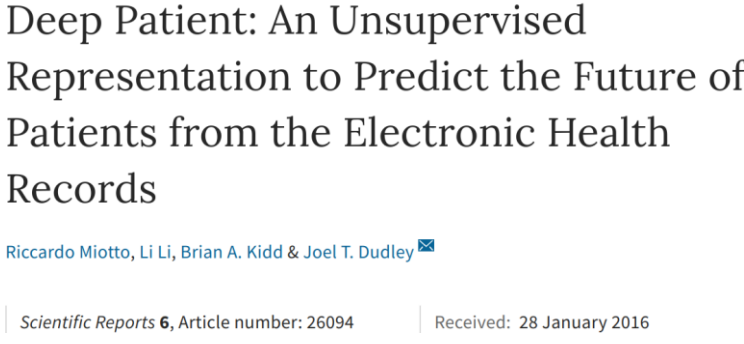
How can we enable 'open ended' hypothesis free research on cohort datasets?

Enable new techniques to find novel patterns in datasets

- e.g. machine learning,

Requires 'open-ended' access+authorisation for research on cohorts

- e.g Deep patient "a novel unsupervised deep feature learning method to derive a general-purpose patient representation from EHR data that facilitates clinical predictive modeling."



Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records

Riccardo Miotto, Li Li, Brian A. Kidd & Joel T. Dudley

Scientific Reports **6**, Article number: 26094 Received: 28 January 2016

Measuring impact

International Cohorts Summit

Global Genomic Medicine Collaborative

- Tracking and quantifying searches
- Measuring wider reuse or impact of particular cohorts
  - Demonstrate value of becoming cohort of cohorts

Examples of how to do this

- Use of DOI to track queries
- Can be used to quantify amount of citations;
- Potential use of distributed ledger to make information publicly available and easier to track