Global Genomic
Medicine Collaborative

International Cohorts Summit

Breakout Group 1

Creating a standardized database and registry – pros, cons, and how best to do it

Breakout Group 1: Creating a standardized database and registry

International Cohorts Summit

Global Genomic
Medicine Collaborative

Our mandate (as we understood it):

How to create a **registry of cohorts** primarily focused on sharing **cohort-level metadata**

If there's time, explore the much more challenging possibility of working towards a **unified database of individual-level data** across cohorts

Breakout Group 1: Creating a standardized database and registry

<u>Discussion question 1:</u> What data would we like to see in a standardized registry of cohorts?

<u>Use cases:</u>

- As a researcher, I want more data to study how an exposure relates to an outcome or replicate a finding.
- As the principal investigator of a cohort, I want to participate in the registry with a minimal amount of administrative burden.

The group proposes a tiered structure

- Surface what is available in each cohort, promoting collaboration
- Where possible, leverage existing metadata collections/questions rather than duplicating effort
- Each cohort contributes up to the tier they're comfortable with
- Queryable (if data are structured)

Breakout Group 1: Creating a standardized database and registry

Tier 0 – Cohort Website

Tier 1 – Cohort Description

- Ascertainment scheme and study design
- Data types and biospecimens available
- Consent (share complete consent form plus structured data use restrictions)
- Data sharing protocol
- Mechanisms and requirements for gaining access to data

Tier 2 – Data Description

- Demographics
- Data collection instruments (exposures, outcomes, etc.)
- ‘Omic data

Tier 3 - Counts

- Tables of counts (e.g., # with particular phenotype, # with urine samples)
- Could be updated periodically (1x/year, 1x/2 years)

Tier 4 – Individual-level Data

Breakout Group 1: Creating a standardized database and registry

International Cohorts Summit

Global Genomic Medicine Collaborative

Discussion question 2: What challenges need to be addressed to optimize the value of sharing information?

- **Fielding inquiries, doing analyses to evaluate feasibility, sharing datasets – all take real work and are not free**
  - Participants were **very** emphatic about this point
  - Where will these resources come from?
  - Could this process be automated?
- Data sharing models will vary by cohort (e.g., 1:1 relationship building vs. limited interaction)
- Infrastructure may require considerable technical work
  - Technology companies may be able to help here, but data should not be held centrally by these companies.
  - Governance needs to be global, independent
- Language challenges – how to deal with translation
- Should not duplicate similar ongoing efforts

Breakout Group 1: Creating a standardized database and registry

Discussion question 3: How might we move towards standardization and sharing of **individual-level data** across cohorts?

- Storing individual-level data poses far greater challenges than storing cohort metadata, including hard limits on data leaving particular countries
- Particular difficulties of phenotype harmonization
- Federation and cloud-based storage seem like the best option

Breakout Group 1: Creating a standardized database and registry

<u>Final discussion:</u> What types of projects could this registry facilitate?

- create a “reproducibility network” to rapidly validate associations discovered in a single cohort
- focus on rare diseases or rare exposures that require massive sample sizes to study
- studies that take advantage of genetic and environmental diversity across cohorts