wellcome connecting science

connecting people with science

Dr Jia-Wern Pan

Data sharing and data management

October 2025

Data Sharing

Outline

- Things to consider
- Practical options for data sharing
- Cancer Research Malaysia’s data sharing policy
- Data management

CRMY Data - Summary

- 7000 patients, 2000 controls
  - BRCA, OVCA, OSCC, NPC
  - Clinical + Epidemiological + Survival metadata
- 6000 SNP genotyping + germline panel sequencing
- 1000 WES Tumour + WES Normal + RNASeq
- 500 sets of H&E + IHC digital pathology images
- 10000 mobile phone images of oral lesions

Data Sharing: Things to consider

- Balancing scientific contributions vs capacity building
  - Facilitating scientific progress
  - LMIC considerations – resource limitations
  - Being “scooped”
- Maximizing impact: citations are not clinical impact
- Commercial interests
  - IP considerations
  - AI training
  - Will your local population be able to benefit?
- Passive data providers vs active collaborators

Data Sharing: Practical considerations

- What to share?
  - Raw fastqs vs mapped bams vs vcfs
- Privacy considerations
  - All individual-level sequence data is potentially identifiable
  - De-identification is not foolproof
- Ethics considerations
  - Is data sharing for future research explicitly approved?
  - What use cases are covered under your ethics approvals?
- How to share:
  - VPN, FTP, SFTP, Cloud storage
  - Publicly hosted databases: GEO, EGA, dbGAP
- Data storage and transfer costs

Data Sharing: Models for data sharing

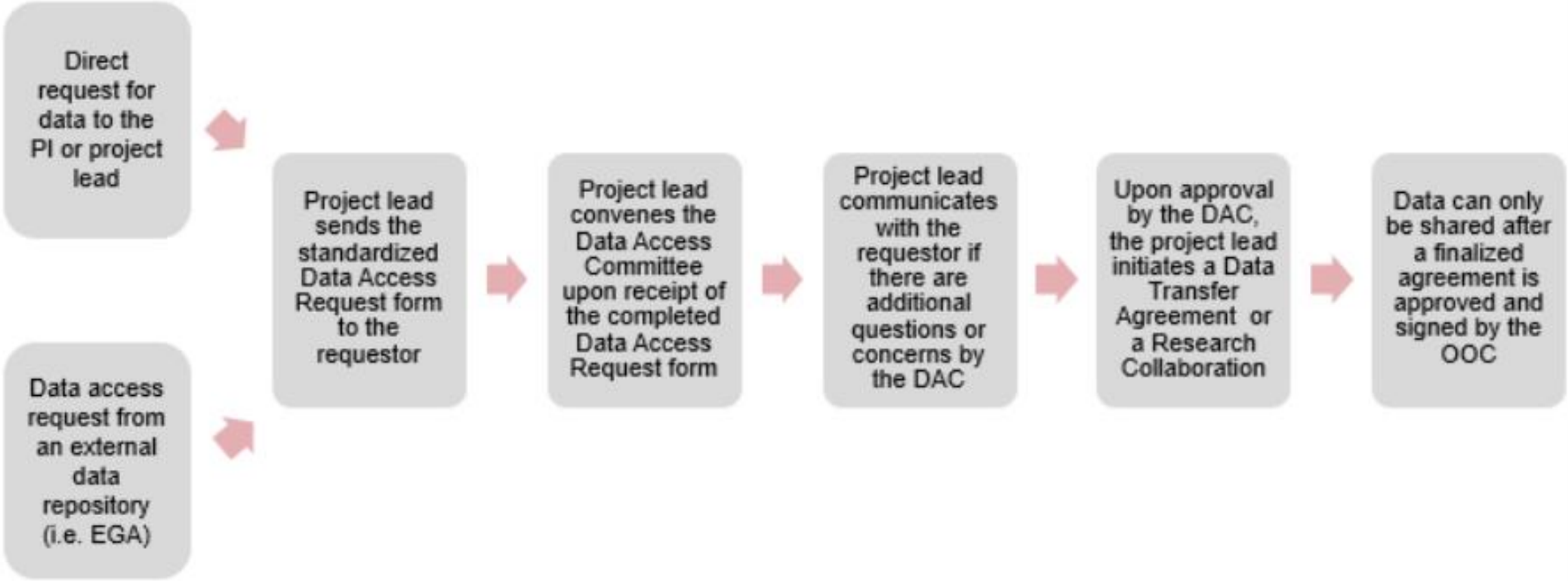
- Fully open access [GEO, 1000Genomes]
- Mostly open access with nominal barriers (registration, data use declaration) [TCGA]
- Moderately open access (DTA, DAC approval required)
- Limited access (Collaboration required)
- Internal use only

Data Sharing Policy at CRMY

Key principles

- (i) **Controlled Access**: CRMY employs a controlled access approach for data sharing requests by external parties.
- (ii) **No Compromise:** No data should be released that would compromise an ongoing trial or study.
- (iii) **Clear Rationale:** There must be a clear scientific or other legitimate rationale for the data to be used for the requested purpose.
- (iv) **Legal and Ethical Compliance:** Data exchange must comply with institutional, legal, and ethical regulations in all the relevant countries.
- (v) **Collaborative Model:** To the extent that is possible, CRMY will share data under a collaborative model that provides opportunities for CRMY scientists to lead on specific analyses within the proposed project. CRMY strongly prefers to share data as active scientific collaborators rather than as passive data providers.
- (vi) **Processed Data Preference:** To the extent that is possible, CRMY prefers to share processed rather than raw data. This means that the processing of CRMY raw data (i.e. with a specific bioinformatics pipeline) should generally be done by CRMY scientists locally, with the processed output data shared instead of the raw data, unless the processing cannot be done locally due to resource limitations.

Data Sharing Policy at CRMY



Assessing Data Access Requests

RATIONALE, MERIT AND CONDUCT OF PROJECT

- What is the scientific merit of the proposal?
- Are the study data suitable for answering the proposed research question?
- What biases might be present? E.g. cohort selection
- Do the original researchers already have plans to use the data in the way proposed by the applicants?
- Are the analyses sufficiently well described to allow assessment of whether the proposal is fit for purpose?
- Is the team properly motivated and suitably qualified to perform the analyses?

PROTECTION OF PARTICIPANTS

- How would confidentiality be maintained?
- Does patient consent cover the proposal? The original ethics approval and consent forms to ensure the proposed use is covered.

Assessing Data Access Requests

PROTECTION OF THE INTEGRITY OF THE ORIGINAL STUDY

- Would the proposed use of data jeopardise the conduct or results of the study from which the data are derived? If so, this is likely to be unacceptable.
- Is there duplication? Have these data already been requested for this purpose? Are there implications for the future conduct or interpretation of the trial?

RESOURCE IMPLICATIONS

- What resources would be required at CRMY to (1) help investigators to understand the data, (2) prepare dataset, (3) transfer dataset, (4) perform analyses?
- Are resources available at CRMY? If not, could they be made available?
- What would be the opportunity costs?
- Are CRMY staff considered providers of data or full collaborators?
- Are raw or summary data required?
- What version of the data set is required? Are new data chasing efforts needed?

Data Management

Key considerations:

- Accessibility vs security
- Data backups vs cost
- Cloud storage vs internal servers
- Ease of access and analysis vs data corruption
- Privacy concerns

Data Management at CRMY

- Epidemiological data/metadata
  - Data manager
  - GCP for anybody working with identifiable data
  - Microsoft Access database
- Genomics summary/aggregate data
  - Internal cBioPortal implementation
  - VPN access
- Genomics raw sequencing data
  - EGA controlled data access
  - Main copies on an internal storage server
  - Backups of raw fastqs on AWS Deep Glacier