MOHCCN Data Access and Use Policy_V1

Table of Content

1. Introduction

The Marathon of Hope Cancer Centres' Network (MOHCCN; the Network) Data Governance Framework supports the responsible sharing of Network data. It is comprised of policies and procedures that support the development of collaborative research, federated learning and harmonize data quality, sharing and access across the Network. Additional policies in the framework include data sharing, data standards, publication, and privacy (Appendix A).

As a key policy within the MOHCCN Data Governance Framework, the Data Access and Use Policy defines the procedures and timelines for access to and use of data shared with the Network (Network Data) in research and publication endeavors. Network Data includes all data that is generated with the use of MOHCCN funding. Data generated with matched funding is encouraged to be made available by the Network and its Partners. Network Data may include genomic and other molecular data types, along with clinical and other health-related information.

Note: Within the initial agreement phase, MOHCCN covers MOHCCN Network members. The next phase of the agreement will include users outside of the Network such as commercial "for profit" users. While this policy is inclusive or different user types, certain access provisions may need to be reconsidered as subsequent phases of the MOHCCN agreement are established.

2. Data Sharing *(Note: timelines to be kept under review)*

The MOHCCN is comprised of individual cancer research projects from across Canada; where clinical information is usually collected at the project site and data, which includes whole genome and transcriptome data, is generated at one of the sequencing centres. Genomic and other molecular data is generated for each specimen, and clinical and other health-related data is collected throughout the timeline of the project. The timelines for sharing should be appropriately set at project set-up to ensure there is enough time for proper quality control and to check the data integrity (Figure 1).

This section covers the use of Network Data from data generation to access for MOHCCN and other users. These policies are in place to ensure the fair and equitable access to Network Data for all researchers to effect the maximal possible insight and discovery and facilitate federated learning.

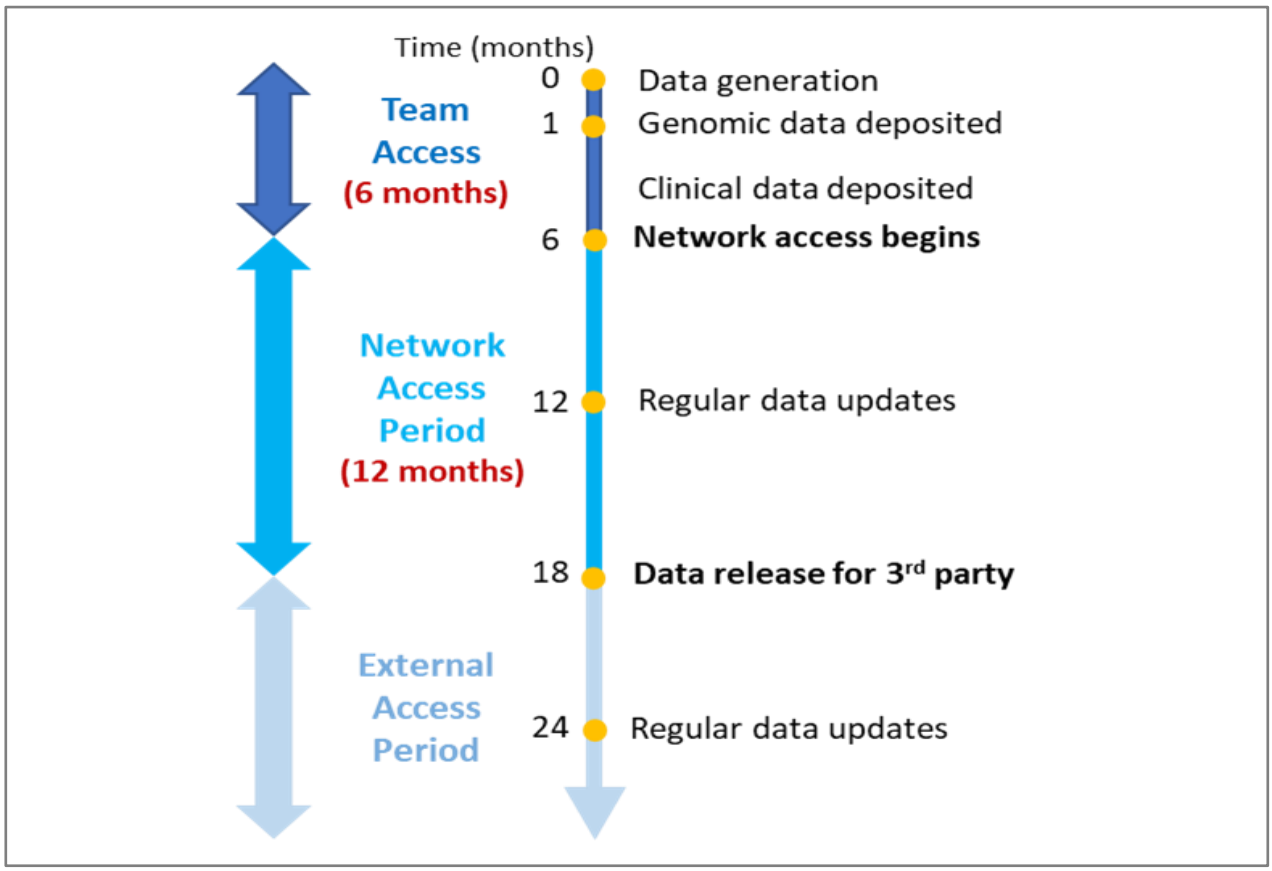

Figure 1

3. Data generation & deposit (1-6 Months)

Network Data are made available to be shared (only under an REB approved study) on a patient-by-patient basis as the data is generated; the data unit outlined in this policy is for individual patients, not datasets or full cohorts.

- **Material Derived Data**

"Material Derived Data" means research subject-level or patient-level data generated from the analysis of Biospecimens, for example, molecular data such as sequencing data, single nucleotide polymorphisms (SNPs), whole genome sequencing (WGS) and whole exome sequencing (WES).

The sequencing centres should share genomic data with the Network in a timely fashion. As a general rule, genomic data for each sample should be deposited and available for sharing within 1 month of completing quality control (QC).

- **Other Molecular Data**

Molecular Data other than Genomic Data may include a wide range of proteomic, metabolomic, cellular and immuno-histochemistry data (timelines to be determined).

- **Clinical Data**

Clinical Data is any medical data that is collected from the patient's medical chart or from the patient and may include patient status, diagnosis, treatment, outcomes, imaging, medical notes, reports and laboratory values. MOHCCN recognizes that the collection of the clinical data is a long-term endeavor, that project sites will need to share data with the Network at multiple time points, and the resources for collection may vary from site to site. Harmonization of data collection will be a priority for the Network.

The mandatory MOHCCN clinical data fields (see Appendix B for MOHCCN Clinical Data/Metadata Standards Guidance) should be made available to the Network as soon as possible, ideally at the time of genomic data sharing, and must be complete by the end of the Team Access period at the very latest (6 months from data generation - see section below). Thereafter, clinical information should be updated for purposes of making such data available at minimum yearly for each case or more frequently if possible. MOHCCN recognizes that some fields will not be applicable to certain cancer types and some historical information, such as prior response to therapies or diagnostic reports, might not be available for collection.

Please find an initial MOHCCN Data Dictionary at this link.

Note: The Data Dictionary is considered a draft and is subject to change as based on the experience of the MOHCCN and approval by the MOHCCN Network Council.

- **Patients on clinical trials**

For patients enrolled in ongoing clinical trials it is acknowledged that certain clinical outcome information pertaining to trial will not be able to be shared until the trial has completed. However, it is expected that clinical information up until the time of enrollment into a trial will be shared. Therefore, some of the clinical fields may be restricted, such as I^{ry} and II^{ry} endpoints and therapeutic treatment resulting from the trial may therefore be withheld beyond 6 months until trial completion or termination. All other clinical data available at the time of biopsy must be shared within the standard timeframe.

The participation in a clinical trial will therefore not exclude the genomic and transcriptomic data from being shared. It should also not preclude any of the clinical data prior to commencement of trial from being collected and provided. The participation of the patient in a clinical trial may be omitted from the provided clinical information until trial completion if preferred by the submitting group. The submitting group should also confirm that they will provide the relevant trial and post-trial clinical information once the trial is completed in a timely manner.

In cases where the trial sponsor agreement and/or pre-existing study requirements precludes or limits the ability to share the genomic data and/or non-trial related clinical information or where intellectual property encumbrances might exist, the submitting group should seek clarification and approval from the TFRI MoH executive group that the generation of genomic data is beneficial to the MoH network or whether the generation of such data should be delayed until trial completion.

- **Other Health-Related Data**

Other Health-Related Data may include patient-related administrative data and patient-reported outcomes data (timelines to be determined).

4. Team Access - Data generation (6 months)

Once the data is available on the Network, it is available only to the team having "contributed" the data for their research purposes for the first six months from the time of data generation (contributed team has ongoing data access). Any data shared later than at six months from generation will immediately be available to all Network members (see below).

5. Network Access (6-18 months)

During the Network access period, Network members may use the data for approved publications and IP development. Network members are expected to work collaboratively and must be respectful and inclusive of their colleagues. By bringing together researchers and clinicians with shared research interests and questions, the Network aims to foster an open and collaborative scientific environment that all participants benefit from.

Any publications resulting from the use of Network resources must adhere to the MOHCCN Publication Policy and Code of Conduct.

- **Data updates**

Projects are required to follow up with their patients regarding any return of individual research results as appropriate, and to provide regular data updates to the Network (e.g., longitudinal clinical data).

- **Analysis & Publications Data**

Publications resulting from the use of Network resources must follow the MOHCCN Publication Policy.

6. Data release (18 months)

At 18 months, non-MOHCCN member researchers in Canada and abroad may now request access to restricted MOHCCN data through the platform *via* a Registered Access or Controlled Access data model. Requests for controlled access data will be made to the Data Access Committee and will require a Data Access Agreement to be in place between the MOHCCN and the third-party researcher and their organization. Data permissions will be uniform within a group and individuals may register by showing proof of membership to the requested group.

Note: In the initial phase, MOHCCN only covers MOHCCN Network members. The next phase will include users outside of the Network and commercial "for profit" users. As these users are onboarded, access provisions may need to be considered.

7. Data Access Model

All Network Data will be shared and available for access by other Network Members, as well as by the wider Canadian and international research communities, following the phases and timelines described above. Furthermore, access procedures and approvals will be contingent on the level of privacy risk engendered by data sharing and use outlined in the Privacy Policy. Unrestricted Access Data will be publicly available for use without any access controls. Network Data requiring access controls will be available to the research community *via* Registered Access and Controlled Access procedures (see Appendix B).

It is important to reiterate, that genomic data that contains information that could potentially identify a participant would not be made publicly available. This includes sequence information and analysis that contain germline variants. Such identifying data would include all sequences deriving from tumour, normal tissue control and transcriptome. Potentially identifying data would always be considered to be protected and restricted. The sharing of protected data would be under the purview and control of the TFRI MOH Data Access Compliance Office (DACO).

8. Addressing Regional Inequity

MOHCCN recognizes that data sharing can be disadvantageous for investigators and centres who do not have the capacity or resources for rapid analysis of their data. However, data sharing can redress an inequity ensuring that a patient's contribution can be studied to its maximal benefit regardless of the capacity of scientists to study it in their local region.

The collaborative Network research process is additionally meant to encourage smaller centres to take advantage of resources that may only be available at larger centres, such as expertise in novel bioinformatic or AI tools. In particular, the first scheduled publication for MOHCCN resources will either be led by, or include, the teams having contributed those materials if they wish to participate in that research. A strong community of collaboration will provide new training and learning opportunities for members of all the Network labs.

Appendix A – DG Framework

MoHCCN Data Governance Framework

- overview of Framework policies -

The Marathon of Hope Cancer Centres' Network Data Governance Framework establishes Network policies and associated procedures to harmonize data quality, sharing and access across the Network.

Policies

- Data Sharing Policy - consent, ethics review, return of results
- Data Access & Use Policy - sharing, accessing and using Network Data
- Publication Policy - authorship and acknowledgment
- Data Privacy Policy - data access models & tiers (Open, Registered, Controlled)
- Data Standards Policy - data models and data/analysis standards and ontologies

Appendix B – DRAFT MOHCCN Clinical Data/Metadata Standards Guidance

DRAFT MOHCCN Clinical Data/Metadata Standards Guidance

The Marathon of Hope Cancer Centres' Network (MoHCCN) aims to generate a dynamic collection of interoperable federated repositories, analytical services, and interactive portals that will allow data to be queried, aggregated, analyzed, and visualized in powerful ways by researchers and clinicians. Access to accurate and comprehensive clinical data related to the patient, their diagnosis and their therapeutic management, is key to enabling discoveries on the MoHCCN platform. The current lack of agreed-upon data models, vocabularies, and ontologies adversely affects interoperability, integration, and analysis across multiple datasets, projects, and repositories. One of the primary goals of the Marathon of Hope Cancer Centres Network is therefore to provide a common data dictionary and tooling that promotes interoperability.

As work to prepare cohorts for sharing and analysis across the Network is underway, the Data Policy & Standards Committee (DPSC) recommends that the **ICGC-ARGO Data Dictionary** be used to describe clinical and other health-related fields and to constrain the contents of the fields. Early experience and feedback from Network sites, along with further definition of MoHCCN data requirements, will inform the ongoing development of the MoHCCN Data Dictionary as well as the data standards roadmap to be developed by the Network.

MOHCCN (mCODE ICGC-ARGO) Data Dictionary

The mCODE initiative[1], led by ASCO, includes both a list of data elements as well as specifications for the coding systems to be used for the value sets to annotate the information in the data elements. With respect to value sets, the mCODE schema uses several international data standards (e.g., LOINC, ICD-10, SNOMED, etc) that can readily be adopted for specific data elements for the Network.

The International Cancer Genome Consortium Accelerating Research in Genomic Oncology (ICGC ARGO) initiative[2] brings together international researchers to analyze genomic and transcriptomic changes along with high-quality clinical data from over 100,000 patients. One of the main goals of ICGC-ARGO is to accelerate the translation of genomic and clinical information into the clinic to guide interventions including diagnosis, treatment, early detection, and prevention.

The ICGC-ARGO Clinical Data Dictionary was developed with the aim of having a uniform set of fields and relationships for reporting patient and tumour samples as well as clinical data concerning lifestyle, environmental exposures, family history of disease, comorbidities, and fields important to precision medicine such as treatment type, treatment response and outcomes, for a broad spectrum of cancers.

The dictionary was developed in consultation with the ICGC-ARGO Tissue & Clinical Annotation Working Group, and adheres to international standards such as ICD-10, ICD-0-3, RxNorm, RECIST, NCIt etc. The MoHCCN Data Working Group conducted surveys across the Network in early 2020 (involving 98 respondents) regarding clinical data availability, ease of gathering and clinical field importance. The results of these surveys and feedback were used to prioritize and further define clinical fields by balancing research utility against data gathering difficulty. The ICGC-ARGO dictionary therefore consists of 86 mandatory clinical fields (64 clinical fields + 22 identifier fields) divided into different clinical entities/events to enable longitudinal clinical data collection.

Alignment of the ICGC-ARGO clinical fields with mCODE and its FHIR implementation indicated good overlap between the two data models (50 fields overlap) and promotes integration and interoperability for projects using mCODE. See detailed mapping: mCODE ARGO table

Links

- **ICGC-ARGO Data Dictionary**
- Mapping between ICGC-ARGO required clinical fields and mCODE v. 1.0 Clinical Fields: mCODE ARGO table
- FAQ

Further development & feedback

We welcome further feedback on the data dictionary as well as on other aspects of data preparation and harmonization.

[1] mCODE: Creating a Set of Standard Data Elements for Oncology EHRs. In: ASCO [Internet]. 4 Mar 2019. Available: https://www.asco.org/practice-guidelines/cancer-care-initiatives/mcode-creating-set-standard-data-elements- oncology-ehrs

[2] https://www.icgc-argo.org/

Appendix C – DRAFT Data Access Tiers

<table>
<tr><td>Open/Unrestricted Access

(Shared via MOHCCN Portal aggregator platforms)</td><td>Registered or Controlled Access MOHCCN Data

(Data accessible to all approved MOHCCN members. Data provided to others with Access approval</td><td>Data Generator/ Hospital</td></tr>
<tr><td colspan="2">Data can be shared with international projects (e.g., ICGC-ARGO) pending MOHCCN agreement in place.</td><td>Data won't be shared</td></tr>
<tr><td>• Limited Clinical Data (e.g., number of patients per disease type, disease type, etc.)
• Somatic mutation data
• Gene/Isoform expression counts
• Derivative data (e.g., annotated somatic)
• Limited germline data (e.g., status of cancer predisposing genes in the germline)
• Pathology Images (de-identified/coded)</td><td>• Detailed Clinical Metadata (e.g., age at diagnosis, sex, etc.)
• Tumour Board & Germline Reports (Targetable Aberrations)
• Treatment & Outcomes
• Fastq and BAM files
• Cancer pathology and imaging information
• As appropriate, Germline genomic information (may require additional approval dependent on DAC)</td><td>• Full date of birth
• Full date of death
• Name
• Provincial healthcare number
• Full dates from clinical data (of hospitals visits, procedures, etc.)</td></tr>
</table>

Document revision history

Developed by	Reviewed by	Endorsed by	Effective Date	Policy Version	Summary of revisions
DPSC	Steering Committee	Network Council	June 2,2022	V1	n/a