MARATHON OF HOPE CANCER CENTRES NETWORK
RÉSEAU DES CENTRES D'ONCOLOGIE DU MARATHON DE L'ESPOIR

MOHCCN DATA SHARING GUIDELINE V1

Table of Contents

1. Introduction

The Marathon of Hope Cancer Centres Network (MOHCCN) aims to create a "Gold Cohort" of 15,000 (15k) cancer cases collected from across Canada over 5 years. The Gold Cohort includes in-depth molecular profiling through whole-genome and transcriptome sequencing (WGTS), combined with rich, standardized clinical information. The Gold Cohort will serve as an invaluable resource to answer scientific questions relating to cancer biology and to accelerate precision medicine for cancer patients.

Network Members (institutions who have signed the Network Master Agreement or a Joinder Agreement to the Network Master Agreement) have agreed to contribute Gold Cohort datasets to a "Network Data Resource". The Network Data Resource is available to Network Investigators (and their designated personnel) for Research Ethics Board (REB)-approved, academic non-commercial data use Studies (after a 6-month embargo period and subject to a transparent controlled-access procedure). The Network has also agreed in principle to make the resource available beyond the Network, including to external (including international) researchers and companies, for both research and intellectual property (IP) development. As described in the MOHCCN Data Access and Use Policy, an 18-month embargo period will apply. Agreements, policies and procedures remain to be defined.

Supported by MOHCCN Committees and Working Groups, the development of MOHCCN policies and guidelines are key components required for the development and responsible sharing of Network data:

- MOHCCN Data Access and Use Policy: Link
- Researcher Code of Conduct: Link
- Clinical Data Model: Link
- MOHCCN Data Privacy Policy: Link
- MOHCCN Data Access Procedures: Link
- MOHCCN Data Publication: Link
- MOHCCN Individual Membership Policy and Procedure: Link

2. Definitions[1]

Authentication: A security mechanism used to verify the identity of a user or device attempting to access a system or application. The CanDIG Platform uses OpenID Connect best practices and KeyCloak tools for authentication (link). These protocols enable each CanDIG Node to authenticate its own individual users using institutional credentials (users log into the CanDIG node at an institution using the same username and password that they use for other institutional resources). Authentication is required before authorized data users can access the CanDIG portal to be able to view and search Network Data.

Canadian Distributed Infrastructure for Genomics (**CanDIG**) **Platform:** A pan-Canadian initiative and platform supporting the distributed management and discovery of genomic and related health data for the Network through the MOHCCN Pathfinder Project. Led by an academic team at University Health Network in Toronto. CanDIG is a technology partner of the

[1] Some definitions are adapted from the Digital Governance Standards Institute, Health Data and Information Lexicon; and the HDRN – Glossary of Terms: Federated Analysis (10 Jan 2024) https://www.hdrn.ca/wp-content/uploads/Federated-Analysis-Glossary-Jan-2024-2.pdf.

DHDP, so it is envisaged that the current data management and sharing activities supported by CanDIG will at a later point be integrated into DHDP. https://www.distributedgenomics.ca/

CanDIG Node: This is an instance of the CanDIG Platform installed at a Network Institution to ingest data so that it is ready to be viewed, searched and accessed. A CanDIG Node comprises the software and protocols that provide the connectivity with other CanDIG Nodes to enable data view, discovery and access.

CanDIG Node Admin: Individual with administrative control over a CanDIG Node who works under the supervision of a Network Researcher. The individual has direct access and control over Network Data in a CanDIG Node, and the ability to make data in a CanDIG Node viewable, discoverable and/or accessible on the CanDIG Platform. A CanDIG Node Admin is responsible to implement an access approval by the Network DAC.

CanDIG Node Data Curator: An individual with authorization to add or edit certain Network Data (e.g., particular Cohort Study datasets) in a CanDIG Node. This role is defined within CanDIG, and an CanDIG Node Admin can add/remove users from the list of Data Curators.

CanDIG Node Institution: A Network Institution with an installed CanDIG Node and that makes Network data available through their CanDIG Node. CanDIG Node Institutions may be a Data-contributing Institution hosting their own data, or an institution designated by a Regional Consortium to host consortium data.

Coded Data: Data where direct identifiers (such as name, date of birth, medical record number, social insurance number) are removed and replaced with a randomly generated number, in order to mitigate risks to individual privacy.

Data Ingestion: Means to move/load Network Data into a CanDIG Node, in a machine-readable format that can support (external) data view, discovery and (eventually) access. Network data ingestion tools and processes, developed by CanDIG through the MOHCCN Pathfinder Project, have been established to provide a mechanism for:

- Validating that clinical from Gold Cohort cases adheres to the MOHCCN Clinical Data Model, and
- Validating that data is stored in a standard machine-readable format and structure.

Digital Health and Discovery Platform (**DHDP**)**:** A TFRI-led platform supporting distributed data sharing and federated learning. The DHDP will incorporate CanDIG as a technical partner and will introduce additional technical components to complement existing CanDIG deployments. The DHDP will support researchers to be able to see a summary of the data available across the Network, to query the data, to ask for permission to access and use data in new precision medicine studies (where approved by REB and Network DAC), and to perform analytic functions across Network Data and other data resources. https://www.dhdp.ca/

Federated Analysis: Analysis of data across multiple datasets, distributed across different locations under the control of different Data Providers, by running the analysis at each location and only sharing and combining the results, in a fast and secure manner.

Federated Learning: A type of Federated Analysis where machine learning is applied to decentralized data. Models are applied locally at each participating institution and only model characteristics (e.g., parameters, gradients) are transferred from the data-holding institution.

Network Cohort Study: A study generating and contributing data to the Network, conducted by a Study Investigator and/or Study Team, generally supported by TFRI funds.

Network Cohort Study Data: Genomic and clinical data generated by a Network Cohort Study and contributed to the Network.

Network Data: Data that meets the Gold Cohort standard and is contributed/shared with the Network. This includes tumor-normal whole-genome and transcriptome sequence data and associated clinical data (as outlined in the most recent Clinical Data Model). Network Data are coded before being contributed to the Network (see "Coded Data" definition).

Network Data-contributing Institution: A Network Institution where data are generated as part of Cohort Studies and contributed to the Network.

Network Data Use Study: An REB-approved academic, non-commercial research Study that has obtained approval from the Network Data to access and analyse Network Data.

Network Data Access Committee (DAC): A Network committee that reviews and approves requests to access Network data for REB-approved Data Use Studies, in compliance with Network policies.

Network Investigator: An investigator who is an individual member of the Network, as defined by the Individual Network Policy. TFRI is responsible for the approval of individual members (under the MOHCCN individual member policy and procedure). Currently, membership (and by extension access approval) is limited to an individual based at a Network Member (signatory of Joinder agreement), bound by the Network Agreement and policies.[2] An individual must also currently be a funded Investigator leading a Cohort Study, nominated Network Scientific Lead, or nominated Co-investigator.

Network Member (also Network Institution): An institution that is a party to the Network Master Agreement (signatory of Joinder agreement), has the associated rights and privileges, and is bound by the terms and conditions of Network Agreements and policies.

Network Pan-Canadian Projects: Multi-site Cohort Studies focused on a common cancer type collaborating to contribute cases to the Network Resources.

Pathfinder Project: An initiative funded by MOHCCN to support CanDIG to develop, pilot and deploy technical tools. The project identifies ways for Network Members to efficiently obtain and curate Network Data and ingest those data into a local CanDIG node. The ultimate goal is to support data viewing, discovery and (eventually) access to Network Data to accelerate precision medicine for cancer patients.

[2] The MOHCCN contemplates expanding data sharing to external researchers after a 18M embargo period (both non-member individuals based at Network Member institutions as well as researchers at external institutions). At this stage, consideration will be needed for expanding view/discovery capabilities to a broader audience.

3. Summary of Network Data Flows

Figure 1 provides a summary schematic describing Network data flows covering:

- Collection of samples, generation of sequence data, and acquisition of clinical data in an electronic data capture system (e.g., REDCap, ATIM).
- Ingestion of Network data into a CanDIG Node (data will initially only be available within the Node until the Node is connected to the rest of the network).
- Connecting multiple CanDIG Nodes to ensure Network Data are viewable and searchable across the Network.

These data flows are expected to be part of MOHCCN's initial data sharing activities, conducted on the CanDIG Platform by an initial subset of Network Institutions and Cohort Study Datasets supported through the MOHCCN Pathfinder Initiative. The Network also envisages that once Network Investigators have identified relevant data on the CanDIG Platform, they will request permission to access and use the data, for REB-approved Studies, from the Network Data Access Committee (DAC). The Digital Health and Discovery Platform (DHDP) will be deployed later to support secure access and federated analysis/learning applications.

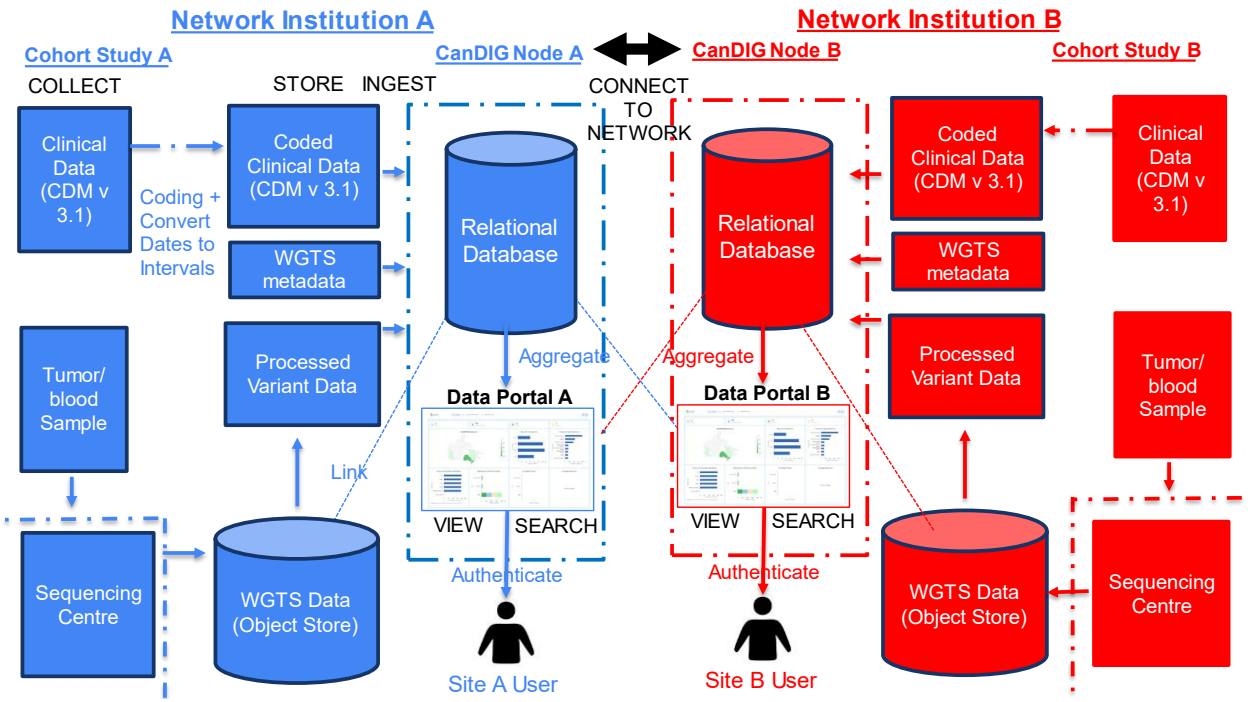


Figure 1. Network Data Collection and Ingestion, View and Search via CanDIG

The Figure shows the connection between 2 Nodes for simplicity. The Network already has plans to adopt 5+ Nodes. Data Portal A will show only summary data from Site A until Nodes are Connected. Once Nodes are connected, users can view summary data from across all connected Sites.

4. Obtaining and Generating Network Data

The aims of the Network are to create and share a Network Data resource consisting of 15,000 cancer cases meeting the Gold Cohort standards (see the Gold Cohort Policy). Gold Cohort data includes tumor-normal whole-genome and transcriptome data alongside associated clinical data (as outlined in the MOHCCN Clinical Data Model).

Data-contributing Institutions send coded biospecimens to regional sequencing centers to generate the WGTS data.

Data-contributing Institutions obtain clinical data either directly from participants or by extracting the data from electronic health records.

Data-contributing Institutions are required to ensure Cohort Study Data conforms with the most recent Clinical Data Model before contribution to the Network.

Data-contributing Institutions are also required to ensure Network data is "coded" meaning that direct identifiers have been removed and replaced with a random, alphanumeric code (see MOHCCN Data Privacy Policy) before contribution to the Network.

Data-contributing Institutions, or when applicable the CanDIG Node, are additionally required to convert any YYYY/MM/DD dates into intervals (time interval between initial diagnosis and event) as part of de-identification, before contribution to the Network. Network tools are available for this conversion from the CanDIG Platform.

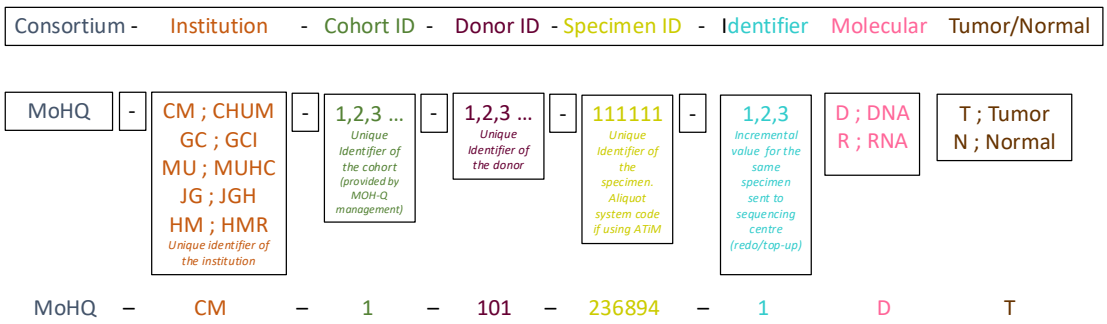
These upfront data standardization, coding and de-identification processes significantly lower the overall privacy risk profile for the rest of the data management and sharing lifecycle, including stages of data storage, ingestion, viewing, search, or access.

Selecting a Unique Patient ID

There is a risk that the same random patient ID, particularly in a multi-site Pan-Canadian project, will be used for different cases ingested at different CanDIG Nodes, because there is no way for each distributed node to check that the identifier is unique across the Network.

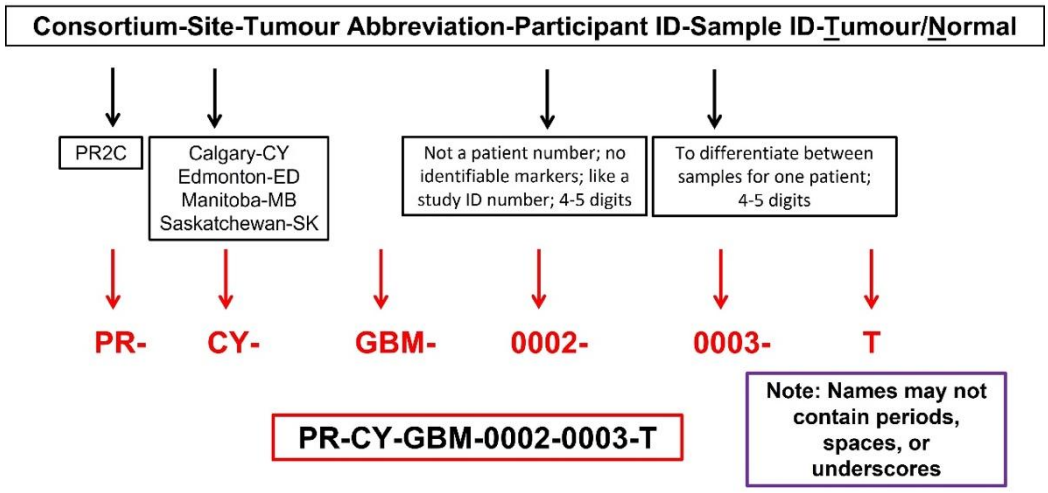
Different Sites should therefore add a standard prefix assigned by the Network at the front of their patient IDs. The prefix should be assigned by the Network in a policy or central registry to ensure that the prefixes themselves are not duplicated.

As an example, the Québec Consortium (MOH-Q) solution is to create a Donor ID that concatenates together an Institution ID (code), a Cohort Study ID (number), and a Patient ID (number). The Cohort Study ID is based on an MOH-Q list, so the same Cohort Study across multiple MOH-Q institutions would have the same Cohort Number. An Excel file to facilitate name attribution has been provided to the cohort study.



MoHQ–CM–1–101–236894–1DT

As another example, the Prairies Consortium (PR2C) add prefixes, including Consortium, Site and Tumor. All PR2C samples are de-identified using the PR2C sample naming convention as depicted. The "Participant ID" is a study-generated ID provided by each cohort and is not an identifiable patient ID. The "Sample ID" is also a study-generated ID provided by each Cohort Study. The naming convention is applied to all genomic, histological, and clinical data associated with each case.



Note: The Network may explore establishing a double-coding mechanism before a future phase where data would be shared with external researchers/companies. This would involve generating a new random code for each sample before granting access to an approved user. This step could be used to reduce the risk of re-identification and to mask the source of the data as appropriate.

5. Distributed Data Deposit and Storage

Cohort Study Teams are required to ingest their Cohort Study data (both genomic and clinical) in a CanDIG Node. A CanDIG Node may be hosted by the same institution hosting the Cohort Study (the data-contribution Network Institution), or by a "central" institution designated by a regional consortium (e.g., MOH-Q). (see **Figure 2**).

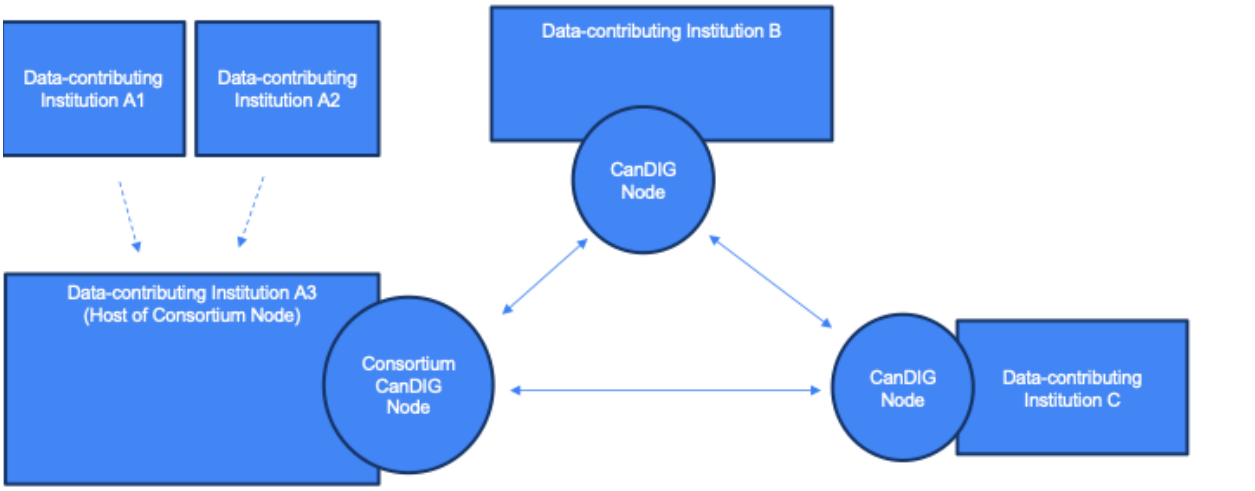


Figure 2. Institutional and Consortium CanDIG Nodes

Dotted arrows represent transfer/deposit of data at another site. Solid arrows indicate sharing aggregate counts and statistics between Nodes.

Each Cohort Study Team is responsible to designate as early as possible the CanDIG Node where its Cohort Data will be deposited and stored.

Where a data transfer between two institutions in a regional consortium is necessary before data can be ingested in a CanDIG Node, the institutions are responsible to establish the necessary legal agreements (e.g., Network Agreement, Consortium Agreement or data transfer agreement) and/or REB approvals to govern the transfer of both the clinical and genomic data.[3] Note that in situations where data are transferred to another location for deposition, storage and ingestion in a CanDIG Node, the CanDIG Node Host Institution will need to be responsible for data security on behalf of the Data-contributing Institution. Moreover, Data-contributing Institutions are responsible to ensure data are properly coded before transfer.

TFRI will work with each CanDIG Node to ensure it has the capacity to securely and sustainably store Network Data (including WGTS data), as well as the technical capability to connect to other CanDIG Nodes in the Network. CanDIG Node Host Institutions must have a general data security framework in place to protect the integrity, confidentiality, and availability of Network Data. They must also have plans in place for long term storage and retention of data (e.g., indefinite, or at least ten years), for the purpose of making data available for future research uses.

Pan-Canadian Projects are generally encouraged to follow the standard policy, with each Data-contributing Institution participating in the project ingesting data in its own CanDIG Node or a consortium CanDIG Node. Pan-Canadian Projects may exceptionally decide to centralize their data storage at one (or more) CanDIG Nodes within the Network, to facilitate

[3] The roles and responsibilities of the initial Data-contributing Institution (i.e., Data Custodian), and the Host Site of the CanDIG Node ingesting, storing and sharing the data, need to be clearly articulated. The Host Site needs to have clearly defined authority (e.g., to share with the Network as authorized by the Network DAC), as well as privacy and security responsibilities. These types of responsibilities are typically laid out in a consortium/collaboration agreement, or in a material/data transfer agreement.

data management, and primary sharing and analysis activities. TFRI works with Projects to identify the appropriate location(s) of their CanDIG Node(s) and transparently report the plan to TFRI. Projects are responsible for any required legal agreements to support data transfers, where applicable (Project Agreement, Data Transfer Agreement).

Restriction on Data Splitting: Splitting of genomic data and clinical data from the same case across different CanDIG Nodes is not currently permitted by the Network or technically supported by the CanDIG Platform. All data for a single case must be ingested into the same CanDIG Node.

6. Data Ingestion

The data ingestion process involves mapping / transforming the clinical data to match the current version of the MOHCCN Clinical Data Model and ingesting that data into the designated CanDIG Node. Ingestion of data into a CanDIG Node ensures the data *is ready* to be viewed, queried and (eventually) accessed by researchers across the Network (and eventually) beyond. Ultimately, the aim is that all of the CanDIG Nodes are technically connected to allow for seamless data viewing, discovery, and (eventually) access across the Network and beyond. For genomic data, this involves 1) connecting local storage holding the genomic data files and 2) ingesting data and metadata to allow for search (this includes experimental metadata about genomic data generation, indexing variant files, and ingesting gene expression matrices). Current step-by-step guidelines for data ingest can be found here. (https://candig.github.io/CanDIGv2/)

Note: Genomic metadata will be defined in a Network Sequence Metadata Policy, including fields such as sequencing machine and library preparation, and will also be made viewable on CanDIG. Transcriptomic data flows are also being defined. It is expected that processed transcriptomic data (e.g., expression counts) will be ingested into relational databases in a CanDIG Node, with raw data kept in the local storage holding other genomic data files. These data flows will be added to future versions of this guideline.

Privacy and Security Safeguards

Sharing data with the Network: There are two steps to making summary statistics and data searches available across the Network. First, data must be ingested into a CanDIG Node. Then, a CanDIG Node Admin must add configuration information to connect ("federate") the CanDIG Node to the other Nodes in the Network. This can be configured in both directions – adding configuration to Site A to be able to display data from Site B, and adding configuration from Site B to be able to display data from Site A. The network sharing is under the control of the CanDIG Node Admin, noting that the sharing is currently all-or-nothing: once connected, a CanDIG Node cannot choose to make only some of its data available for summary statistics and data searching.

Privacy: The ingestion process supports privacy protection, as it provides validation that only data meeting the current MOHCCN Clinical Data Model controlled fields are included in the dataset that will potentially be made viewable and discoverable across the Network.

7. Viewing Summary-Level Data on CanDIG Portal

CanDIG Nodes display summaries of their deposited/ingested Network Data on the Node's Summary Page (see **Figure 3**). Summaries include descriptive statistics of what data is available in the CanDIG node and connected CanDIG nodes across the Network. Only authenticated individuals at each CanDIG Node are able to view the Summary Page hosted by that CanDIG Node.

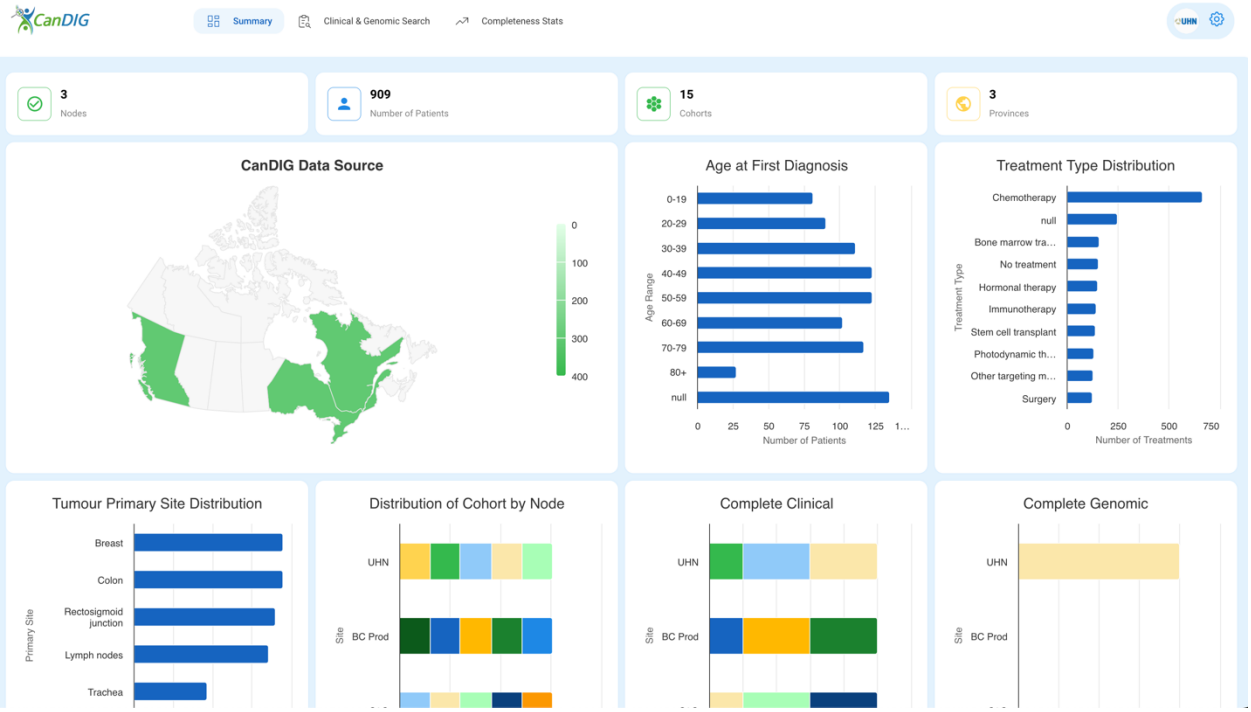


Figure 3. CanDIG summary page

Showing aggregated data across the network (synthetic data shown)

Publication of data summaries is an important objective of the Network to provide information to researchers about what data is available. This enables Network Investigators and their designated personnel to find relevant data, develop study and collaboration ideas, and prepare a data access request to send to the Network DAC for an REB-approved data use Study.

CanDIG tools and protocols can generate standard, summary data from select clinical data and sequence metadata fields and make these available on the Summary Page. The Network objective/standard is that CanDIG Nodes publish the following summaries of their Network Data on the CanDIG Portal:

- Descriptions of the Cohorts.
- Aggregate statistics and completeness statistics generated from selected clinical data fields. The selected clinical fields used to generate summary statistics are illustrated in **Figure 3** and currently include age at first diagnosis, treatment type, tumor primary

site and number of cases per CanDIG Node with complete genomic and/or clinical data.
- No summaries are currently generated from genomic or transcriptomic data.

Privacy and Security Safeguards for Publication of Summary-level Data on the CanDIG Portal

CanDIG Node Approval: Network Data ingested into a CanDIG Node will only be displayed in other CanDIG nodes once connection to the other CanDIG Nodes is configured by the CanDIG Node Admin. In turn, the CanDIG Node Admin will ensure appropriate approvals have been obtained from privacy and security offices before connecting the Node.

Node-level Aggregation Threshold: CanDIG applies a Node-level aggregation threshold when publishing data summaries to the CanDIG Portal. A standard aggregation threshold of "n<10" will be applied across the Network to all viewable summary statistics. Node-level aggregation means that a single site will only return "n<10" if there are less that 10 results. Nodes will return "0" if there are no matching results. In other words, the statistics shown from each CanDIG Node will either be 0, <10, or an integer of 10+. This threshold was established by UHN for piloting the technical platform based on standard guidelines.[4] This has now been implemented as a standard threshold by the CanDIG Platform, to establish a consensus privacy protection baseline as well as to facilitate technical implementation.

Security: The CanDIG Portal Summary Page will initially only be viewable by authenticated Network Investigators and their designated personnel. CanDIG provides protocols for authentication. Each CanDIG Node authenticates its own users accessing its own data portal using institutional credentials. Each CanDIG Node will maintain a whitelist of Investigators, who are individual members of the Network, and the personnel designated by Investigators as needing access to the CanDIG portal.[5]

Note: When MOHCCN seeks to expand access to external researchers/companies, summary data will need to be made viewable to a broader community of potential requestors.

Note: Summaries displayed may change from time to time based on user feedback. Because privacy is protected through aggregation, there is currently no policy restriction on what underlying clinical data fields and sequence metadata fields from the Clinical Data Model

[4] The rationale for this threshold is based on existing Canadian standards for anonymization of clinical trial data. Government of Canada, Public Release of Clinical Information: guidance document (2019) https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html ; Ontario IPC, Deidentification-Guidelines-for-Structured-Data.pdf (ipc.on.ca) (June 2016); Statistics Canada, Disclosure Control, https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/control-controle-eng.htm (archived); Government of Canada, Privacy Implementation Notice 2023-01: De-identification https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information-privacy-notices/2023-01-de-identification.html ; Tri-Council Policy Statement (2022, https://ethics.gc.ca/eng/tcps2-eptc2_2022_chapter5-chapitre5.html . Quebec Regulations on Anonymization (2024) https://www.publicationsduquebec.gouv.qc.ca/fileadmin/gazette/pdf_encrypte/lois_reglements/2024A/106829.pdf

[5] CanDIG Pathfinder is currently developing the technical functionality by April 2025 to control access to specific individuals.

are used to generate data summaries. For technical and user experience reasons, only a subset of fields will be displayed on the summary page.

8. Data Discovery on CanDIG Portal

Data discovery: Network Investigators and their designated personnel, who are authenticated by a CanDIG Node, can submit a search query composed of certain clinical fields or genomic variants (position or gene-based) (see **Figure 4**). The platform will only return aggregate case counts, descriptive statistics and location. Case counts are the number of cases that match the selected features. Descriptive statistics mean a breakdown of the output cohort (the matching cases) by age, treatment type etc., subject to the aggregation threshold described above to limit return of small sample sizes. Case counts will be given for the Network overall, as well as broken down by each CanDIG Node.

The list of fields/data types that can be queried are currently the following and can be expanded or modified through additional user testing: genomic variants (by gene name or position), as well as the following fields from the Clinical Data Model: Treatment type, Tumour Primary Site, and specific drugs for Systematic Therapies.

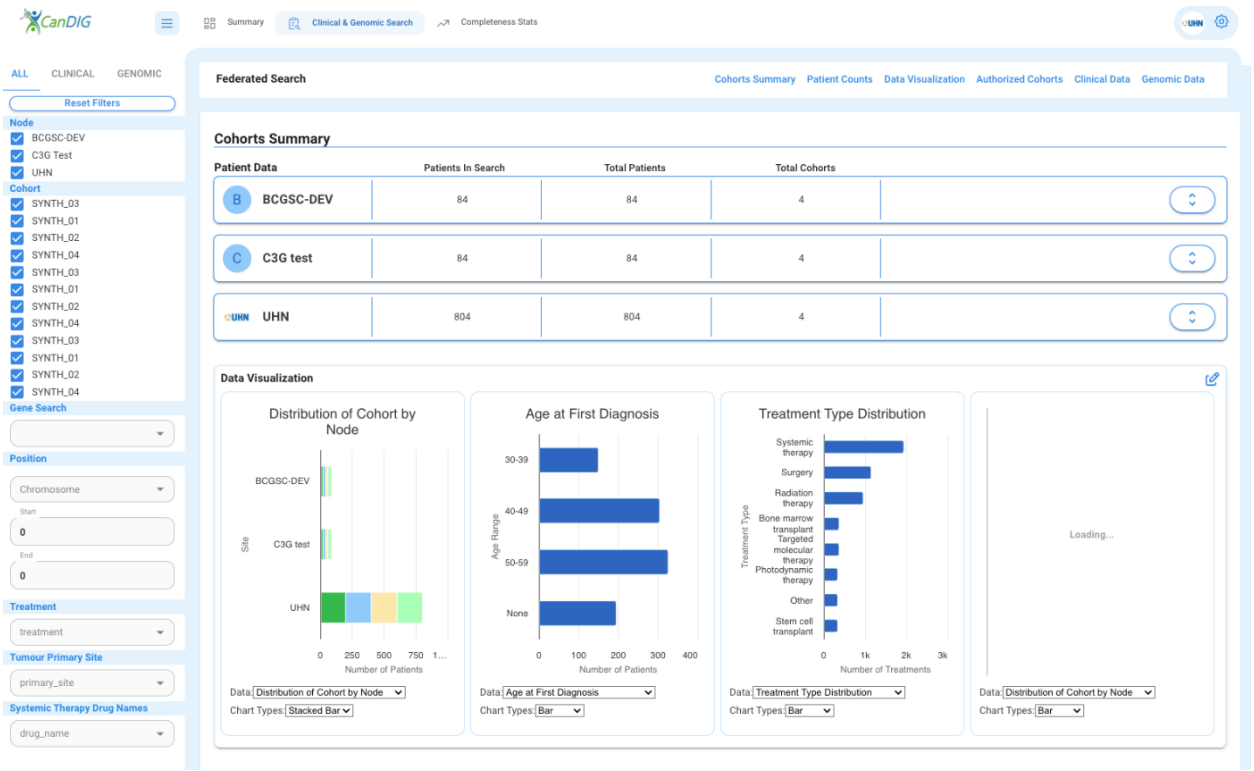


Figure 4. CanDIG Search Page

Showing search fields on sidebar and tabular / graphical results. Showing synthetic data only.

Privacy and Security Safeguards for Data Discovery

CanDIG Node Approval: Network Data ingested into CanDIG will only be displayed on other CanDIG Nodes once connection to the other CanDIG Nodes is configured by the CanDIG Node Admin, after approval by the CanDIG Node Institution (e.g. privacy and security).

Privacy: The CanDIG search protocol includes the same Node-level aggregation safeguard described above for data view (n<10). This ensure that only aggregate case counts are delivered as responses to discovery queries.

Security: The CanDIG search function will initially only be available by authenticated Network Investigated and their designated personnel. CanDIG provides protocols for authentication using institutional credentials.[6]

9. Controlled-Access Procedure

This section summarizes the Network's current controlled-access procedure for access to row-level data (i.e., individual-patient level clinical and genomic data). Note that while the operational procedures for access requests and approvals have been defined, the Network is still defining the technical mechanisms/platforms by which controlled-access data can be securely accessed and analysed by approved users (to be outlined in future versions of the guideline).

In terms of the operational procedures, a Network Investigator must be approved by the Network DAC in order to be granted "access" to row-level data from a requested set of MOHCCN Cohorts, for use in a REB-approved Study (see MOHCCN Data Access Procedures). A Network Investigator must list designated personnel on the data access request who need access to conduct the Study. The Network DAC can only approve access to studies that have obtained REB approval. The Network DAC currently has the authority to approve academic, non-commercial studies that relate to MOHCCN's mission. All data-contributors have agreed and are responsible to ensure that data contributed to the Network Resource are obtained from patients who have provided consent for future research use. The Network DAC therefore assesses requests against general criteria, and not against any specific consent terms.[7]

[6] Additional safeguards that could be considered (where proportionate) include: More detailed audit logs of user actions, e.g. screens viewed, and searches conducted. A definition of security relevant events would be needed to trigger breach detection and response (e.g., excessive # of searches suggestive of a re-identification attack). A basic user agreement (or individual membership agreement) reminding users of DHDP of their responsibilities, e.g., not to export data; to keep credentials secure.

[7] Some (mainly retrospective) Cohorts may be subject to certain limitations on sharing stemming from consents, REB or institutional approvals, or privacy laws. To date, there are no potential data use limitation reported by Cohorts that are relevant to the initial pilot phase of MOHCCN data sharing activities. As regards anticipated future phases of MOHCCN, it has been reported that some consent templates do not (clearly) permit sharing with industry researchers. Similarly, it is also possible that MOHCCN will in the future accept requests to recontact participants for recruitment or additional data collection, or to link data with other databases (e.g., administrative health records). At such a time that MOHCCN extends access to industry researchers for scientific research and/or commercial use, the Network would need to offer Cohorts a means to communicate and enforce this data use limitation. The GA4GH Data Use Ontology will be implemented to permit Cohorts to exceptionally tag Cohort data with data use limitations (e.g., non-commercial use) with reasonable justification. This will require additional technical development and data management processes.

Note: Currently the Network's governance framework only supports data access by Network Investigators (and their designated personnel) for REB-approved, academic, non-commercial data use Studies. Additional agreements, policies and procedures will be needed before other types of requestors and uses are permitted.

Network access will be requested, approved, and implemented on a Cohort-by-Cohort basis, i.e., to Study Cohort datasets selected and justified as relevant for the proposed Study. Access can be granted to all Study Cohort datasets for a proposed data use Study where justified (e.g., pan-cancer analysis).[8] Requestors can use the data view/discovery functionality on CanDIG to identify Cohorts containing cases with clinical or genetic characteristics of interest. Data access requests will list the Cohorts of interest on the data access request form, as part of the form sent to the Data Access Committee secretariat by email.

Approved Network Investigators are bound by the Network Agreement – Terms of Access and Use, requiring them to only use data for the approved Study and to protect privacy, security and confidentiality. The Network DAC and Data-Contributing Institution(s) shall track duration of access and terminate access (or request data destruction/return) after 2 years (with 30-day notification and potential for renewal). The Data Recipient is responsible to destroy the data if the project is completed or terminated.

In terms of the technical means of access and analysis, the ambition of the Network is to make controlled-access data available to approved users on the secure DHDP for federated analysis/learning. The Data Policy and Standards Committee recommends that download/transfer between Network Members also be supported for approved Studies led by Network Investigators. This is the traditional means of data sharing in genomics and health research, where physical copies of data are transferred to the user to analyze in their own computing environment. For a Study involving download/transfer of data, additional DAC procedures would need to be developed, requiring additional assurances that the user's computing environment is secure, and that the user would be held responsible for maintaining security. As part of the Pathfinder Project, CanDIG is developing a technical mechanism to support secure data transfer/download (to be detailed in future versions of the guidelines).[9]

To date, the Network is only aware of limitations on data sharing at the Cohort level (stemming from the consent template) rather than the patient level (stemming from specific patient choices on an individual consent form). (In theory, granular choices could be offered/made by patients relating to data sharing that would need to be managed, but a concrete use case has not yet been reported in the Network. If participants have opted out of data sharing all together, they must not be included in the MOHCCN Gold Cohort.) The current technical proposal is to enable data use limitations to be applied at the Cohort level only.

[8] Functionality in CanDIG to implement user permissions for specific cohorts is technically ready to go (API-only – currently there is no graphical user interface for site admins to issue or change data access).

[9] Some examples of data download mechanisms used by the community are described here. Globus is used for the BQC19 data. Access is given to Globus endpoints. Data can be partitioned in different ways. One possibility could be to create an endpoint per MOH cohort and bundle clinical and genomic data there. This kind of manual solution requires human resources to work properly (data administrators managing access, maintaining the endpoints, etc.). Another currently available mechanism is for researchers to submit study Cohorts to the European Genome-Phenome Archive, which currently supports the transfer of genomic and related health data to approved users.

Document Revision History

Developed by	Reviewed by	Endorsed by	Effective Date	Policy Version
Data Sharing Subgroup	Steering Committee	Network Council	February 28, 2025	1

Authors

Name	Institution	Title
Karen Cranston (Chair)	UHN	Technical Project Manager
Adrian Thorogood (Chair)	TFRI	Legal Advisor and Data Governance Manager
Jeffrey Bruce	UHN	Scientific Associate
David Bujold	McGill	Bioinformatics Manager
Christine Caron	CHUM	Project Manager
Eric Chuah	BCGSC	Group Leader, Bioinformatics - Databases
Daniel Gaston	Dalhousie	Lead, Bioinformatician
Paul Gordon	U of C	Bioinformatics Manager, CHGI
Nicolas Luc	CHUM	Manager, ATiM
Valez Lumi	AHS	Senior Advisor
Jessica Nelson	BCC / BCGSC	Projects Team Leader
Véronique Ouellet	CHUM	Project Manager
Carolyn-Ann Robinson	U of C	Senior Research Associate
Enrique Sanz-Garcia	UHN	Assistant Professor / Staff Medical Oncologist
Marion Shadbolt	UHN	Project Manager
Lillian Siu	UHN	Senior Scientist
Emily Van de Laar	UHN	Research Associate II