

Evaluating the problem of finding and accessing restricted data in Canada

Where do we go from here?

Land acknowledgement

I live and work on Treaty 6 Territory and the Homeland of the Métis. I pay my respects to the First Nations and Métis ancestors of this place and commit to being an ally to and partner with those who came before me.



<https://indigenous.usask.ca/index.php>

Restricted data?

Data that are not immediately accessible immediately accessible because they are restricted or only available upon request.

Examples:

- Licensed data (e.g. commercial, organizational)
- Restricted data (e.g. health, legal)

Example



The screenshot shows the CLSA website's Data Access section. A yellow arrow points to the 'Data Access Application Process' link in the sidebar. The main content area is titled 'Data Access Application Process' and contains information about the COVID-19 Questionnaire Study data availability, application requirements, and contact information for requesting a Magnolia user account. A large blue button labeled 'Apply for CLSA Data' is prominently displayed at the bottom of the main content area.

Data Access

- DataPreview Portal
- Data Access Application Process**
- Data Access Resources
- Application Deadlines
- Fees
- Data and Biospecimens
- FAQs

Data Access Application Process

CLSA COVID-19 Questionnaire Study data are now available.

Please note that all requests for access to the COVID-19 Questionnaire Study data must be made through a NEW data access application in Magnolia. The usual data access fees will apply, even if you are a current approved user. Requests cannot be submitted through an amendment for an existing application.

Please email access@clsa-elcv.ca to request a Magnolia user account, providing your full name, institutional email, position title and institution as part of your request. Additional information on trainee access is available [here](#).

Apply for CLSA Data

We strongly advise potential applicants to review all of the information on this page before applying to use the CLSA platform.

The Discovery/Access Problem

Where's the data?

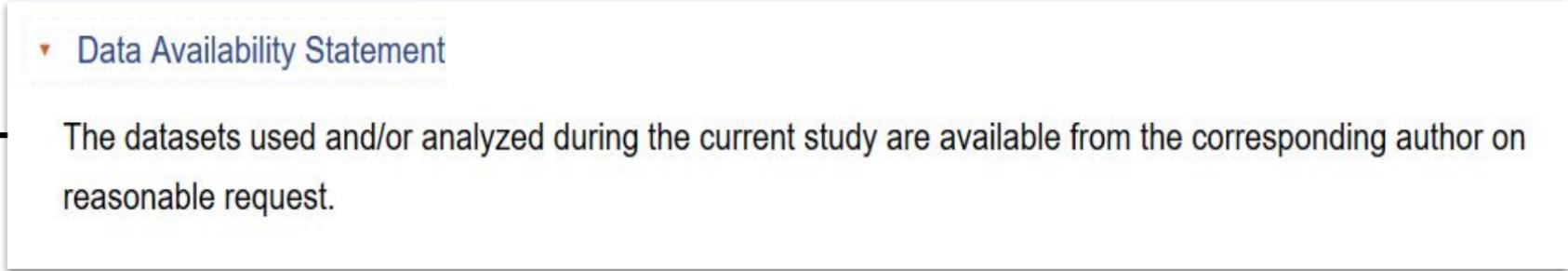
Research says:

- The ability to search for restricted data through individual data sources/websites is difficult or non-existent
- Researchers struggle to discern whether or not restricted data can be used for research purposes
- Researchers are often unaware that restricted data exists at all

Where's the data?




Issue is exacerbated when researchers report on their use of restricted data:



▼ Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

"None of the publications that required an application included metadata sufficiently outlining the requirements for access and approval."



Read KB, Ganshorn H, Rutley S, Scott DR. Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis. Canadian Medical Association Open Access Journal. 2021;9(4):E980–7. <https://doi.org/10.9778/cmajo.20200303>

Accessing the data...

"the re-use of this [restricted] data requires a set of complex approvals from multiple governing entities which are often opaque, difficult to navigate and obtain, and so pose risks to population based research"

Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. *Int J Popul Data Sci.* 2018;3(3):432.

- Uncertainty about whether they are eligible to access the data
 - Difficulties both in understanding and navigating the application request process
 - Lack of standardization of how to submit a data request across sources
 - Amount of time it takes to apply for and acquire restricted data is prohibitive and an impediment to research
 - Many data sources do not provide adequate support to help someone navigate the process
-

Understanding the data...



Restricted data is often of poor quality (e.g., not maintained, poorly organized, lacks standardization, data quality is low)

Documentation necessary to facilitate reuse is often absent or insufficient

These barriers have consequences



Researchers limit their research questions to data they can easily find and obtain

Researchers may invest a substantial amount of resources into acquiring data that cannot be easily acquired and/or used

Academic and non-academic research is limited when restricted data does not provide means to make data easily discoverable and/or accessible

Why does this matter?

The Canadian data sharing landscape does not currently support the discovery and access of restricted data in the same way it does open data

Restricted data can have tremendous value (e.g. health outcomes) but is not optimized for discovery, accessibility, and reuse

Addressing the Problem

Forming a national working group

Access Limited Data Discovery Working Group operating within the Digital Research Alliance of Canada

Access-Limited Data definition broader than “restricted data”

Goals:

1. Scope the landscape of Canadian access-limited data locations, platforms and/or tools;
2. Identify the challenges associated with increasing the discovery of access-limited data;
3. Make recommendations for improving the discovery of access-limited data based on challenges identified.

Working group membership

- Kevin Read, University of Saskatchewan (Chair)
- Grant Gibson, Canadian Research Data Centre Network
- Amber Leahey, Scholars Portal
- Lynn Peterson, National Research Council
- Sarah Rutley, University of Saskatchewan
- Julie Shi, University of Toronto iSchool (Graduate Student Assistant)
- Victoria Smith, Digital Research Alliance of Canada
- Kelly Stathis, DataCite
- Jeremy Geelen, Canadian Institutes of Health Research (Special Advisor)

Research questions

RQ1: What types of Canadian access-limited data sources exist that include datasets that could be used for research purposes?

RQ2: How well do a sample of Canadian restricted health data sources identified in RQ1 make their data discoverable and accessible?

RQ3: What are the challenges associated with discovering and accessing restricted data from the sample of Canadian health data sources reviewed in RQ2?

Our Approach

Step 1: Scoping Canadian data sources

Identify access-limited Canadian data sources

Reviewed:

- Canadian data sources (e.g. CRDCN, Scholars Portal)
- University/college and academic partnership websites
- Government websites

Requested submissions from Alliance RDM
Expert Groups and CANLIBDATA listserv

Info captured:

Geographic Region

Sector

Disciplinary Focus

Canadian access-limited data source inventory

n=137

Data Source Name	Region	Data Source Description copy and paste from data source	URL	Date URL Last Accessed	Associated Institution(s)	Sector(s)	Discipline Use CRDC classification
Abbvie Pharmaceutical Research & Development	National; International	There are important health benefits in making clinical trial data and information available to health care providers, researchers, patients and the general public. Thus, we have adopted national and international principles and standards regarding the sharing and publication of clinical trials data and information. In addition, AbbVie conducts audits of compliance to ensure we are meeting those principles and standards.	Link	2022-02-18	Abbvie (private company, operates globally but headquartered in the US)	Private	Medical, health and life sciences
Agriculture and Resource Development, Government of Manitoba	Manitoba	The Agriculture and Resource Development Midland Sample and Core Library is a secured library and viewing facility of Manitoba cores and samples collected from wells drilled under The Oil and Gas Act, the Manitoba Stratigraphic Core Hole Program, and select mineral exploration drillcore collected through the Mines Act. There is no charge to view the cores in the facility. Informational Notice 18-02 states our policies on examining, sampling, analyzing and removing these core and samples that were collected through the Oil and Gas Act.	Link	2022-02-18	Government of Manitoba	Government	Engineering and technology
Alberta Health Services	Alberta	<i>Not provided</i>	Link	2022-02-18	Alberta Health Services	Government	Medical, health and life sciences
altalis	Alberta	We are committed to the continued updating, maintenance, storage, distribution and licensing of Alberta's primary spatial datasets and our trusted partner's imagery and LiDAR datasets.	Link	2022-02-18	Various (multi-partner initiative)	Non-profit; Government; Private	Other
Association of Faculties of Medicine of Canada data	National	40 years worth of data on Canada's medical education system	Link	2022-02-18	AFMC	Other	Medical, health and life sciences
Atlantic Canada Conservation Data Centre	Atlantic	AC CDC maintains location data for species of conservation concern (S-ranks between S1 and S3S4, in addition to species considered extirpated [SX] and historic [SH]) in a Geographic Information System (GIS) database.	Link	2022-02-18	Atlantic Canada Conservation Data Centre	Non-profit	Natural sciences
Atlantic PATH	Atlantic	Atlantic PATH has recruited over 34,000 participants from all four Atlantic Provinces. The samples and information that participants have given will help researchers find out why some people develop certain chronic diseases and others don't. This information will help to find new ways of preventing chronic diseases and to diagnose these diseases earlier, when they can be easier to treat.	Link	2022-02-18	Dalhousie University & Canadian Partnership for Tomorrow's Health	Academic	Medical, health and life sciences

Step 2: Grading Canadian health data sources

Identified **55 health data sources** in our inventory

48 were eligible for review

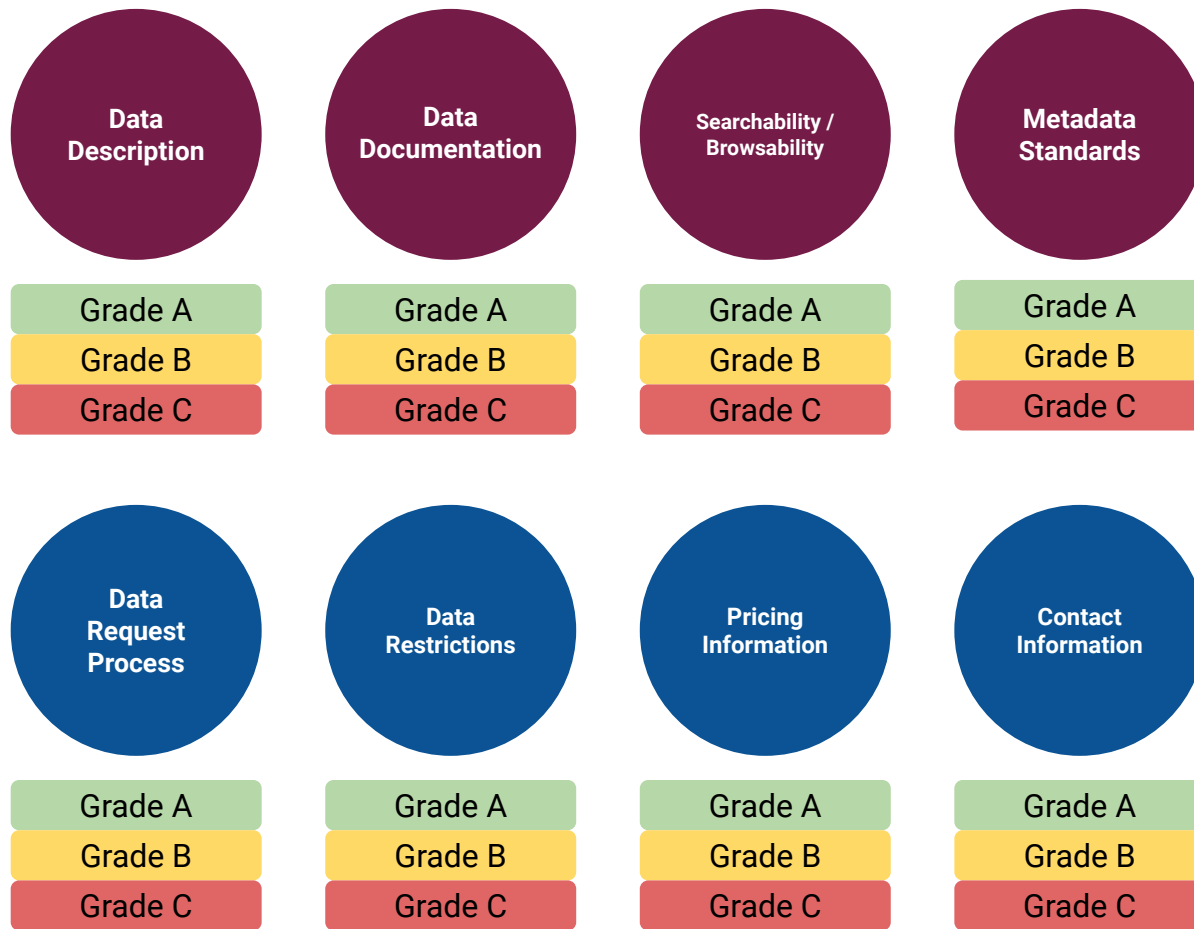
- n=4 only permitted patient data requests
- n=3 became inaccessible during review

Each source underwent qualitative review to identify discovery/access attributes

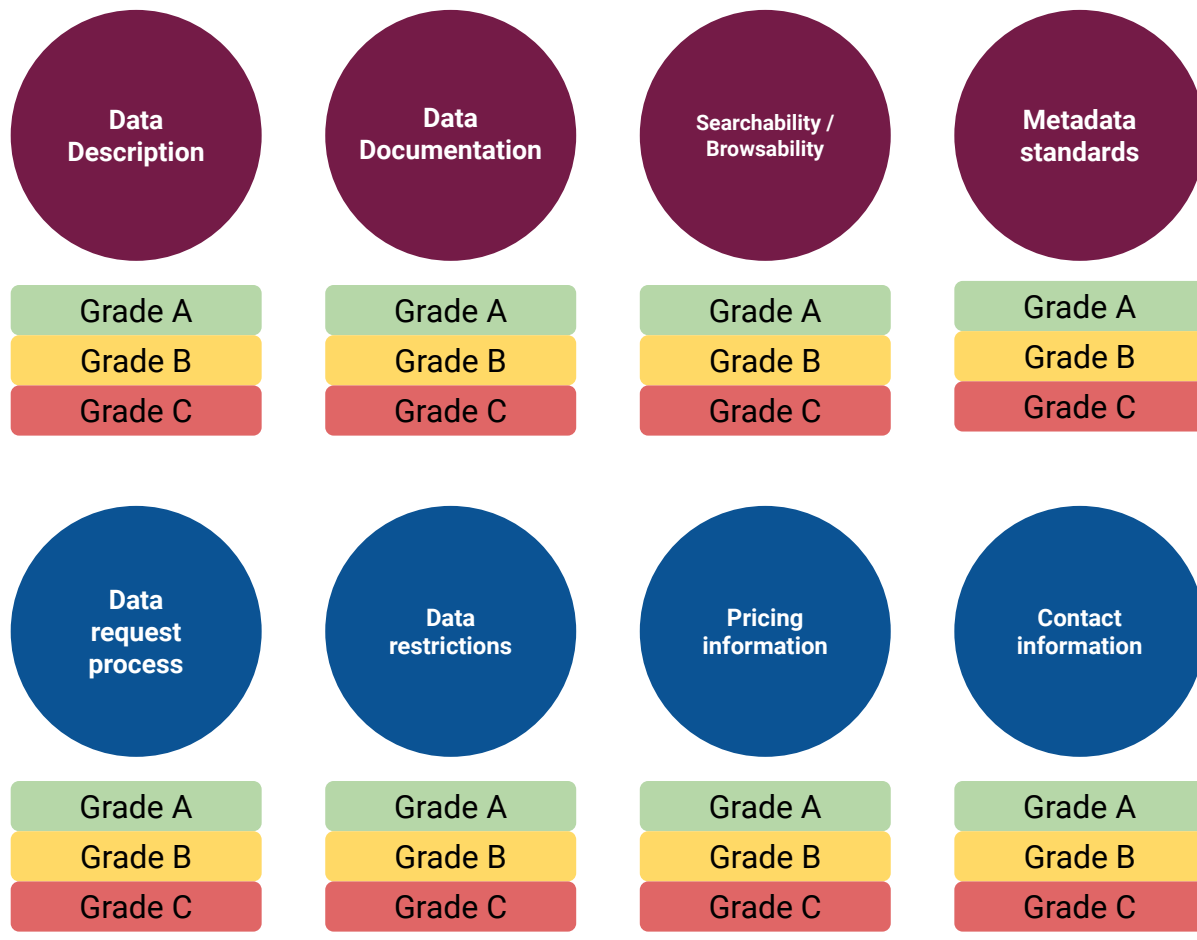
Attributes Identified:

Discovery Attributes	Access Attributes
Data description	Data request processes
Data documentation	Data restrictions
Searchability / browsability	Pricing information
Use of metadata standards	Contact information

Attribute	Definition
Data description	Description of the data itself, including summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and major research questions.
Data documentation	Detailed, structured information about the data itself that supports its interpretation and use.
Searchability /Browsability	If a data source has more than one dataset, there is the presence of mechanisms for searching and browsing them (e.g. full search interface, table, list, etc.)
Use of metadata standards	The presence of metadata standards applied within the data source. Metadata standards establish a common way of structuring data about the dataset. For data sources that do not have individual datasets within them, this section can be applied to the data source as a whole.
Data request processes	The presence, completeness, and clarity of information and/or content (e.g. application forms) required in order to successfully submit a data access request and to understand the data access request process
Data restrictions	The criteria of persons/organizations/projects who are eligible to access the data. Data restrictions should inform a user whether or not they would be eligible to acquire the data, and should indicate what parameters are necessary to meet eligibility.
Pricing information	Descriptions of transparent pricing information on data source websites.
Contact information	Information provided that clearly describes how to get in touch with someone from the organization in the event that they have a question about either the data or the request process?



Grading 48 health data sources



- 2 authors reviewed each data source
- Tiebreaker for discrepancies

Grading goals

Evaluate the state of discovery and access in Canadian health data sources

Identify gaps and areas for improvement

Prioritize data sources for future indexing (e.g., FRDR, Dataverse)

Inform final recommendations

Grade A

Grade B

Grade C

Discovery grading criteria



Data Description

Grade A

1. Provides description of the data that supports understanding for a broad audience
2. Describes who is responsible for creating the data
3. Describes the data's intended use or purpose

Grade B

1. Description of the data requires more info to facilitate understanding and selection for reuse
2. Description of who is responsible may not be entirely clear
3. Description of data's intended use or purpose may not facilitate understanding, selection, or reuse

Grade C

1. Little or no detail about the data,
2. Little or no detail about who is responsible for the data
3. Little or no detail about the data's intended use or purpose

Discovery grading criteria



Data Documentation

Grade A

1. Datasets are accompanied by multiple pieces of detailed, structured information about the data itself, as well as additional contextualizing information. Interpretation and reuse of the data is possible with the materials provided.
 - e.g. data dictionary, codebook, user guide, code, data collection materials or instruments

Grade B

1. Datasets are accompanied by some detailed, structured information about the data itself.
 - Data dictionary, codebook, user guide, or robust data collection instrument is provided, but more may be required to facilitate interpretation and reuse

Grade C

1. Datasets are not accompanied by information about their content and structure. Interpretation and reuse of the data is not possible with the available materials.

Discovery grading criteria



Searchability / Browsability

Grade A

1. Datasets are searchable within the data source
2. Basic keyword search is available
3. There is a mechanism to browse datasets by one or more relevant facets or variables (e.g., topic, population)
4. There is an advanced interface allowing boolean search and/or searching of specific metadata fields

Grade B

1. Datasets are searchable within the data source
2. Basic keyword search is available

Grade C

1. There is no mechanism to search or browse the data source

Discovery grading criteria



Metadata Standards

Grade A

1. Structured metadata is available and structured using one or more recognizable standards (e.g. Dublin Core, ISO 19115)
2. Metadata elements are employed consistently. There is a set of core elements, including metadata needed to cite the dataset
3. Metadata is clearly presented within the web interface

Grade B

1. Structured metadata is available, but may not adhere to a recognizable standard
2. Metadata elements are employed consistently; there is a set of core elements that is usually present (e.g. title, description)
 - a. Metadata may not be presented in a conventional interface for this tier. For example, a text document with structured fields would meet criteria.

Grade C

1. There is no metadata, or
2. Metadata is present but it is limited and unstructured—e.g. a title and description only

Access grading criteria



Data Request Process

Grade A

1. Source clearly explains all aspects of the request process including application forms, instructions, timelines, review process, and submission process
2. Support for submitting data access requests is available and clearly indicated

Grade B

1. Source clearly explains most aspects of the request process including application forms, restrictions, timeline, and submission process

Grade C

1. Source provides no clear instructions on how to submit a request or what to include, or
2. Data request form is not available

Access grading criteria



Data Restrictions

Grade A

1. Source includes a section devoted to describing data restrictions
2. Specific populations or projects eligible to access data are well described and include examples
 - a. e.g. “Only oncology researchers affiliated with [institution] may access these data”
 - b. e.g. “Only research studies concerning [topic] may use these data”

Grade B

1. Restrictions on who can use the data are mentioned but not described

Grade C

1. No information is provided, or
2. Information is not clear enough to determine if restrictions exist

Access grading criteria



Pricing

Grade A

1. Fees are explained clearly, with enough information to estimate the cost of a specific request
2. If a quote process is in place, a researcher can:
 - a. Submit a basic research plan to receive a quote;
 - b. Contact someone for help estimating fees; or
 - c. View a sample project with fees applied

Grade B

1. It is clear that there are fees for accessing data
2. Estimated costs are provided
3. If a quote process is in place, a full research proposal is required

Grade C

1. No information on fees is provided

Access grading criteria



Contact Information

Grade A

1. There is a clear, easy-to-find contact person and email address devoted to data inquiries and applications
2. Contact info is displayed within the area of the website related to data

Grade B

1. Contact information related to the application process is provided
2. Contact info is displayed within the area of the website related to data

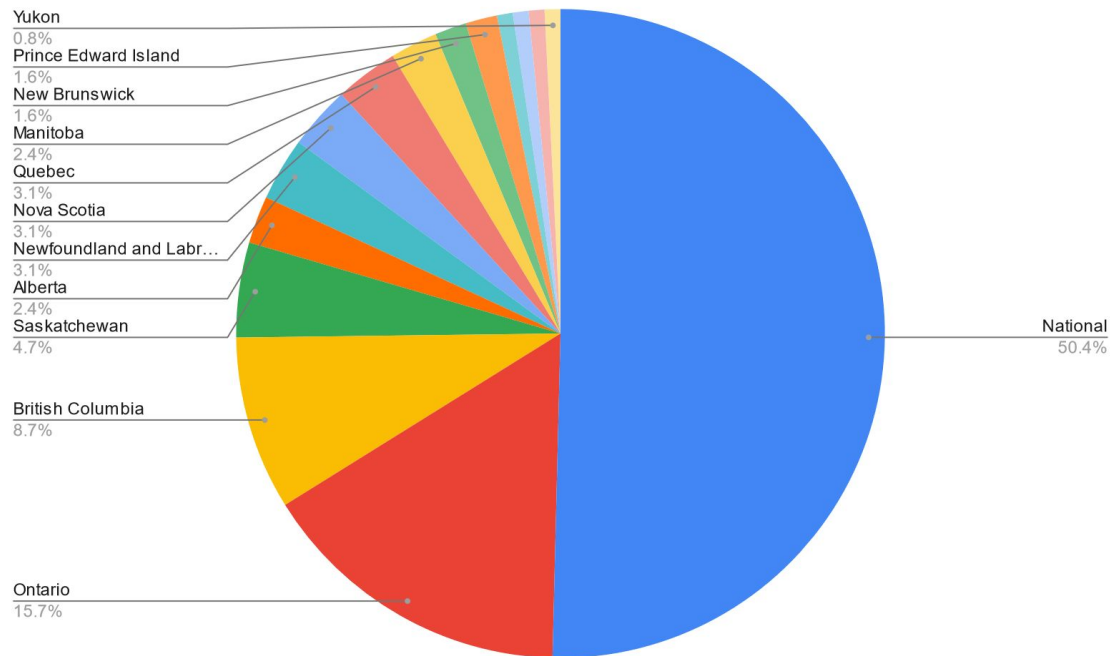
Grade C

1. No contact information is provided, or
2. Only a generic “Contact Us” tool is provided

Results

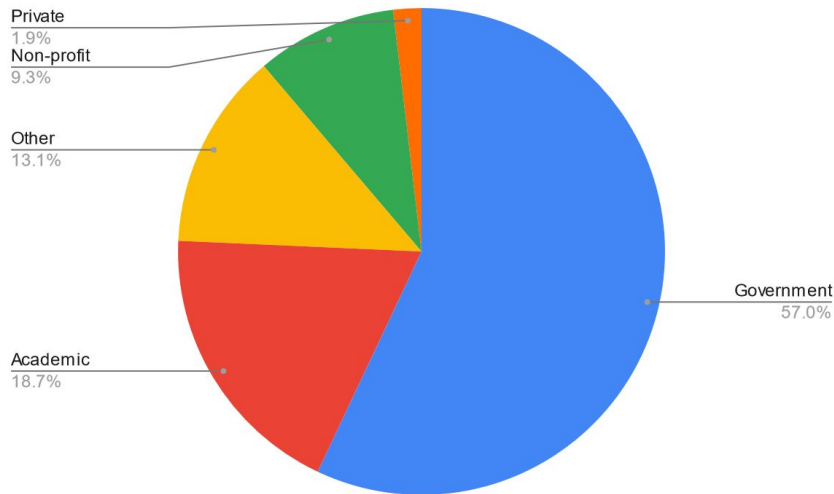
Canadian access-limited data landscape

Region
(n=137)

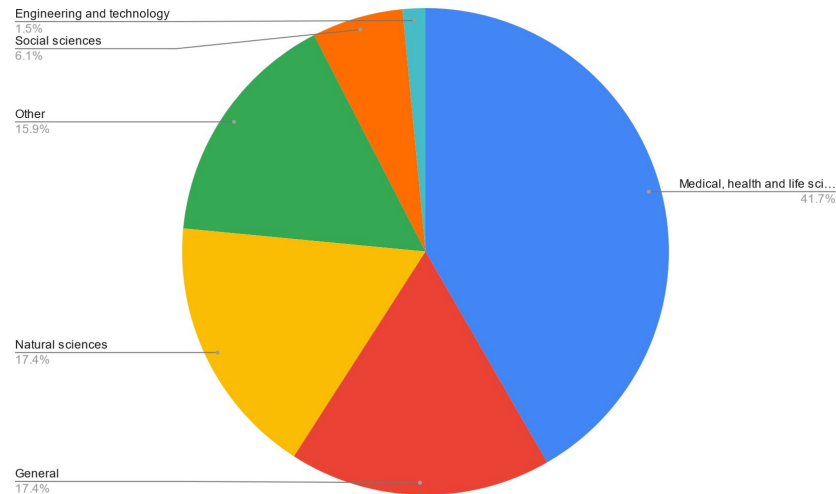


Canadian access-limited data landscape

Sector



Discipline

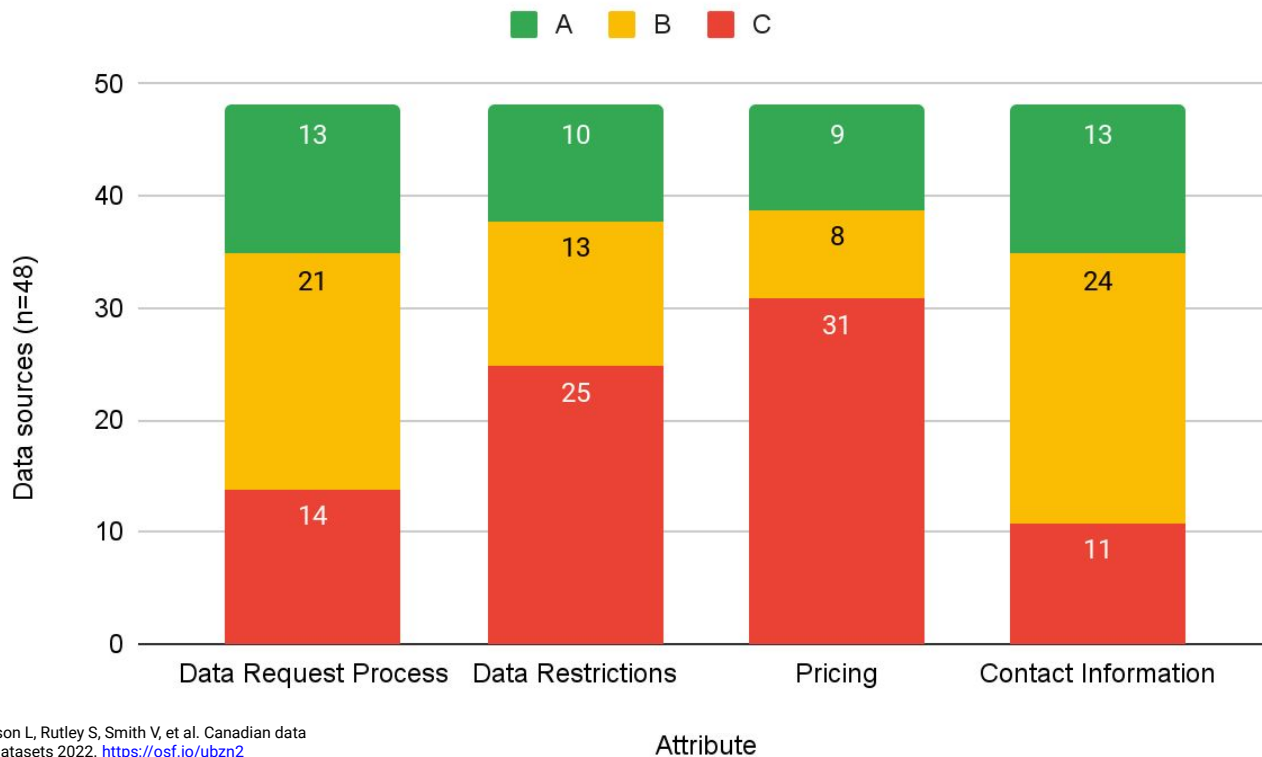


(n=137)

Discovery grading of health data sources



Access grading of health data sources



Key takeaways

- **42% (n=20)** did not receive an “A” grade in any category
- **52% (n=25)** did not provide any information about data restrictions
- **56% (n=27)** did not provide any data documentation to support interoperability and/or reuse
- **79% (n=38)** received a “C” grade for metadata standards
- **0% (n=0)** received an “A” grade for metadata standards

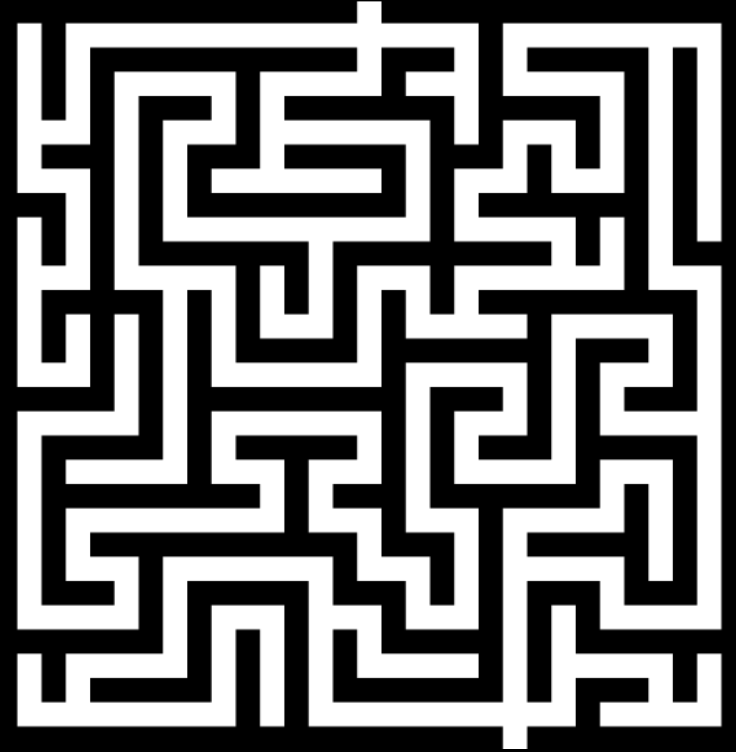
It's not all bad! Potential data sources for indexing

Rank	Data Source	# of A Grades	# of B Grades	# of C Grades
1	Canadian longitudinal study on aging	6	0	1
2	Manitoba Population Research Data Repository	4	4	0
3	CanPath	4	3	1
4	Population Data BC	4	3	1
5	BORN ontario	4	2	2
6	Canadian Institute for Health Information	3	4	1
7	ICES	3	4	1
8	Cancer Care Ontario	3	3	2
9	Health Data Nova Scotia	3	3	2
10	BC Cancer Registry	3	2	3
10	CHILD cohort study	3	2	3

Key Barriers & Recommendations

Insufficient Infrastructure

Barrier 1



Infrastructure barriers



Navigation, workflows, and linkage to related content were challenging

Dataset information and access request information were often separate

Time investment in seeking and learning about datasets was very high

Lack of standardization / high variability in each data source

Some data sources vanished during our study – preservation concern

Infrastructure: Recommendations

1. Establish a community of practice for stewards of Canadian restricted data sources to establish commonly accepted guidelines and standards;
2. Offer funding opportunities for restricted data sources to adopt data discovery and access standards;
3. Align data request procedures as much as possible across jurisdictions to improve workflows for the acquisition of restricted data (e.g., Sensitive Data Pilot); and
4. Explore a common infrastructure model that could be adopted by all restricted data sources in Canada.

Where's the metadata?

Barrier 2

Dataset Description: ?

Population studied: ?

Timeframe: ?

Access procedure: ?

Dataset restrictions: ?

Cost of data: ?

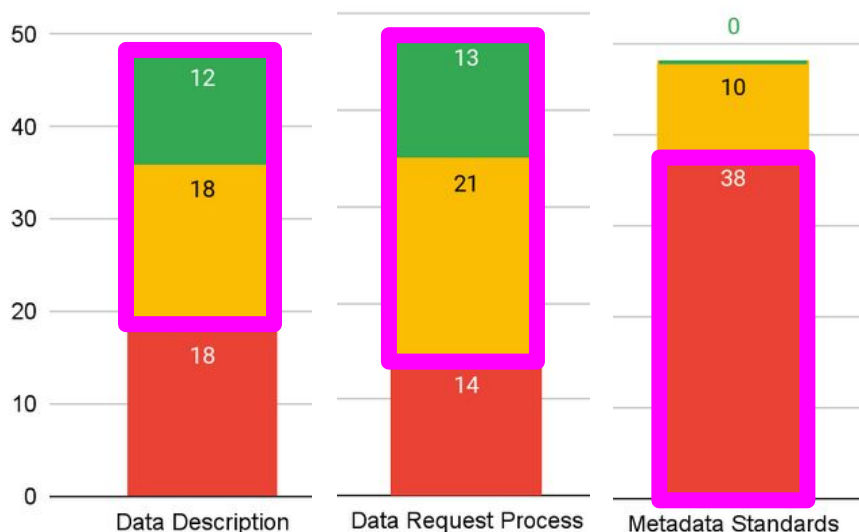
Metadata barriers

Grading identified reasonably good descriptions of data and access procedures, but no metadata to disseminate them;

Non-existent structured metadata for the data access request process;

Dataset metadata schemas provide little specificity in this area; and

Without metadata, these data sources will remain hard to find and their data will not be accessed



Metadata: Recommendations

1. Develop a metadata schema (or extension) that addresses **access-specific** data requirements to improve discovery and access; and
2. Engage with data sources who were identified through our grading exercise to improve the discovery of their data in national aggregators (e.g., FRDR).

Metadata: Recommendations

1. Develop a metadata schema (or extension) that addresses **access-specific** data requirements to improve discovery and access; and
2. Engage with data sources who were identified through our grading exercise to improve the discovery of their data in national aggregators (e.g., FRDR).

Metadata: Imagine the possibilities

Patient Centered Measurement: The Acute Inpatient Survey 2016/17

Data sources

78 Acute Care Hospitals and 2 freestanding rehabilitation hospitals in 6 health authorities (Fraser Health, Interior Health, Island Health, Vancouver Coastal Health, Northern Health, Provincial Health Services Authority) and Providence Health Care.

Date range

September 1 2016 to March 1st 2017

Description

The Acute Inpatient 2016/17 Survey asked patients about their health-related quality of life and their experiences with the quality of care and services received as an inpatient in one of 78 acute care hospitals and two freestanding rehabilitation hospitals in British Columbia.



The survey was coordinated by the BC Office of Patient-Centred Measurement (PCM) on behalf of the BC PCM Working Group, a group that includes representation from the BC Ministry of Health and the seven Health Authorities.

The Survey included items from the following Patient Experience Reported Measures (PREMs) and Patient Outcome Reported Measures (PROMs).

View a video presentation on these data below:

Access criteria:

- Project proposal required
- REB ethics approval required
- Data transfer agreement necessary

Dataset restrictions:

Only researchers with active Tri-agency grant funding are permitted to access this data.

Conditions of use:

Researchers are only permitted to access data on secure systems that have no internet access.

Timeline to access:

~1-2 months processing

Cost of data:

- \$250: Access
- \$40/hour: Support

Lack of Documentation

Barrier 3



Documentation barriers



We identified very little data documentation across health data sources;

Mirrors challenges identified in existing research (e.g., time investment); and

Many data sources do not have the personnel or expertise to develop robust documentation (no devoted data stewardship support)

Documentation: Recommendations

1. Develop targeted guidance and training for restricted data sources to articulate the value, importance, and utility of including data documentation for restricted data; and
2. Engage large organizations with robust administrative staff who have well-documented data to provide guidance for those with less funding/capacity

Concluding Thoughts

Our next steps



Publish our findings from this research

Progressing on phase 2 of our work to explore discovery and access metadata for restricted datasets:

- Extract metadata from 48 health data sources
- Identify commonalities in metadata elements
- Develop a minimal metadata standard/extension for both:
 - Restricted datasets
 - Restricted data access procedures

Going forward

Canada has room for improvement with respect to the discovery and access of restricted data

High value restricted datasets that can be used for research are often hidden, inaccessible, and/or unusable

Metadata for restricted data and their access procedures can improve discovery and reuse

Only scratched the surface with health data sources (82+ data sources remain)

Support from national data initiatives is crucial for future success in this area

References

1. Bekemeier B, Park S, Backonja U, Ornelas I, Turner AM. Data, capacity-building, and training needs to address rural health inequities in the Northwest United States: a qualitative study. *J Am Med Inform Assoc*. 2019;26(8–9):825–34.
2. Boland MR, Karczewski KJ, Tatonetti NP. Ten simple rules to enable multi-site collaborations through data sharing. *PLoS Computational Biology*. 2017;13(1):e1005278.
3. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nature genetics*. 2020;52(7):646–54.
4. Clayton GL, Elliott D, Higgins JPT, Jones HE. Use of external evidence for design and Bayesian analysis of clinical trials: a qualitative study of trialists' views. *Trials*. 2021;22(1):789.
5. Garrison NA, Barton KS, Porter KM, Mai T, Burke W, Carroll SR. Access and Management: Indigenous Perspectives on Genomic Data Sharing. *Ethnicity & disease*. 2019;29(Suppl 3):659–68.
6. Hanna CR, Lemmon E, Ennis H, Jones RJ, Hay J, Halliday R, et al. Creation of the first national linked colorectal cancer dataset in Scotland: prospects for future research and a reflection on lessons learned. *Int J Popul Data Sci*. 2021;6(1):1654.
7. Ho HKK, Gorges M, Portales-Casamar E. Data Access and Usage Practices Across a Cohort of Researchers at a Large Tertiary Pediatric Hospital: Qualitative Survey Study. *JMIR Med Inform*. 2018;6(2):e32.
8. Knosp BM, Craven CK, Dorr DA, Bernstam EV, Campion TR. Understanding enterprise data warehouses to support clinical and translational research: enterprise information technology relationships, data governance, workforce, and cloud computing. *J Am Med Inform Assoc*. 2022 Mar 15;29(4):671–6.
9. Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. *Int J Popul Data Sci*. 2018;3(3):432.
10. Mpango J, Nabukenya J. A Qualitative Study to Examine Approaches used to Manage Data about Health Facilities and their Challenges: A Case of Uganda. *AMIA Annu Symp Proc*. 2019;2019(101209213):1157–66.
11. Prince K, Jones M, Blackwell A, Simpson A, Meakins S, Vuylsteke A. Barriers to the secondary use of data in critical care. *J Intensive Care Soc*. 2018;19(2):127–31.
12. Rahimzadeh V, Schickhardt C, Knoppers BM, Sénécal K, Vears DF, Fernandez CV, et al. Key implications of data sharing in pediatric genomics. *JAMA pediatrics*. 2018;172(5):476–81.
13. Read KB, Ganshorn H, Rutley S, Scott DR. Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis. *Canadian Medical Association Open Access Journal*. 2021;9(4):E980–7.
14. Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. *Health Serv Res*. 2010;45(5 Pt 2):1442–55.
15. Sarwate AD, Plis SM, Turner JA, Arbabshirani MR, Calhoun VD. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. *Frontiers in neuroinformatics*. 2014;8:35.
16. Saulnier KM, Bujold D, Dyke SO, Dupras C, Beck S, Bourque G, et al. Benefits and barriers in the design of harmonized access agreements for international data sharing. *Scientific data*. 2019;6(1):1–6.
17. Simpson CL, Goldenberg AJ, Culverhouse R, Daley D, Igo RP, Jarvik GP, et al. Practical barriers and ethical challenges in genetic data sharing. *International Journal of Environmental Research and Public Health*. 2014;11(8):8383–98.
18. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. *Nature medicine*. 2016;22(5):464–71.
19. Sydes MR, Johnson AL, Meredith SK, Rauchenberger M, South A, Parmar MK. Sharing data from clinical trials: the rationale for a controlled access approach. *Trials*. 2015;16(1):1–6.

Resources

1. Read KB, Gibson GA, Leahey A, Peterson L, Rutley S, Smith V, et al. Access-Limited Data Source Grading Rubric 2022. <https://osf.io/kc4u9>
2. Read KB, Gibson GA, Leahey A, Peterson L, Rutley S, Smith V, et al. Canadian data source identification and evaluation datasets 2022. <https://osf.io/ubzn2>

Questions?



kevin.read@usask.ca