

Improving the Discovery of Restricted Data in Canada

Identifying Metadata Commonalities Across Restricted Data Sources

Digital Research Alliance of Canada:

Access Limited Data Discovery Working Group

Presenter:

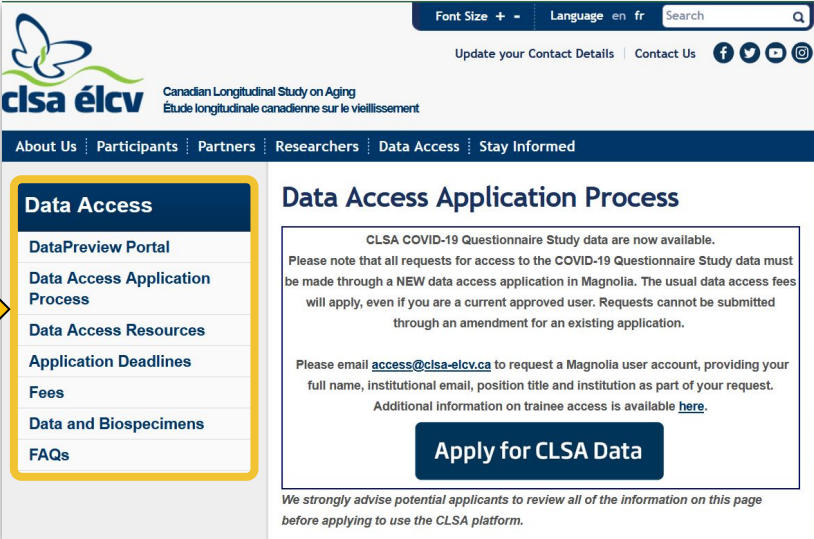
- Kevin Read, University of Saskatchewan (Chair)

Members:

- 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
- Victoria Smith, Digital Research Alliance of Canada
- Kelly Stathis, DataCite

What do we mean by restricted data?

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



The screenshot shows the CLSA website's Data Access page. The header includes the CLSA logo, the text 'Canadian Longitudinal Study on Aging / Étude longitudinale canadienne sur le vieillissement', and navigation links for 'About Us', 'Participants', 'Partners', 'Researchers', 'Data Access', and 'Stay Informed'. A yellow arrow points to the 'Data Access Application Process' link in the left sidebar. The main content area is titled 'Data Access Application Process' and contains the following text:

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.

Restricted Data Discovery & Access Challenges

Where's the data?

Restricted data is hard to find, and even more difficult to access

▼ 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 restricted data is a known issue

Known researcher access challenges:

- Uncertainty about eligibility for access
- Difficulties navigating the request process
- Lack of standardization for submitting data requests
- Time burden
- Lack of support from data provider(s)

These barriers have consequences

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

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

Research is constrained when restricted data cannot be used

Our Project:

Identifying and Evaluating
Restricted Data Sources in Canada

Project goals

1. **Find** (as many) examples of Canadian restricted data sources
2. **Evaluate** Canadian restricted data sources based on how well they make their data discoverable and accessible
3. **Extract** metadata commonalities from restricted data sources to test alignment with existing metadata schemas

Project goals

1. **Find** (as many) examples of Canadian restricted data sources
2. **Evaluate** Canadian restricted data sources based on how well they make their data discoverable and accessible
3. **Extract** metadata commonalities from restricted data sources to test alignment with existing metadata schemas

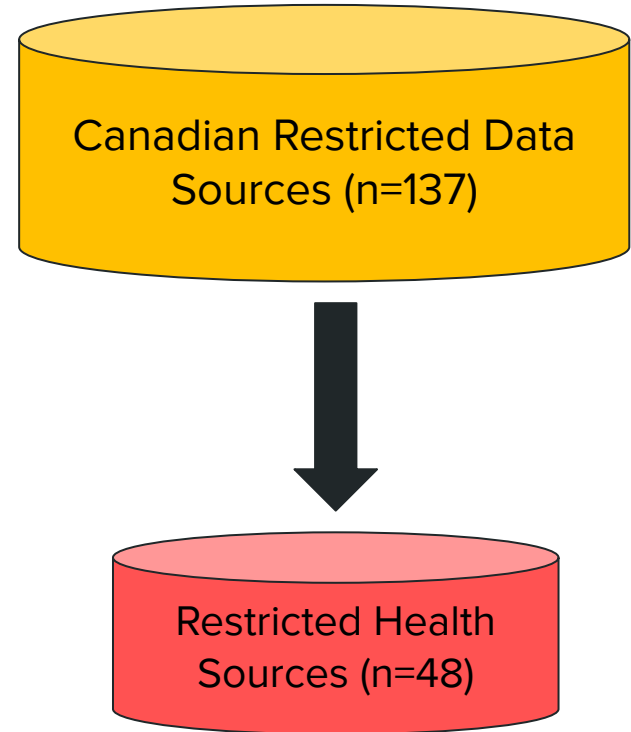
Canadian restricted data sources

Search Strategy:

- Academic websites
- Government websites
- Private sector
- Non-profit sector
- Web-search
- Call to experts

Health sources were most prominent (n=48)

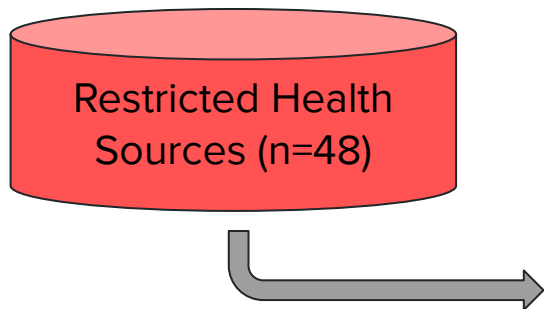
- Used health sources as sample for remainder of project



Project goals

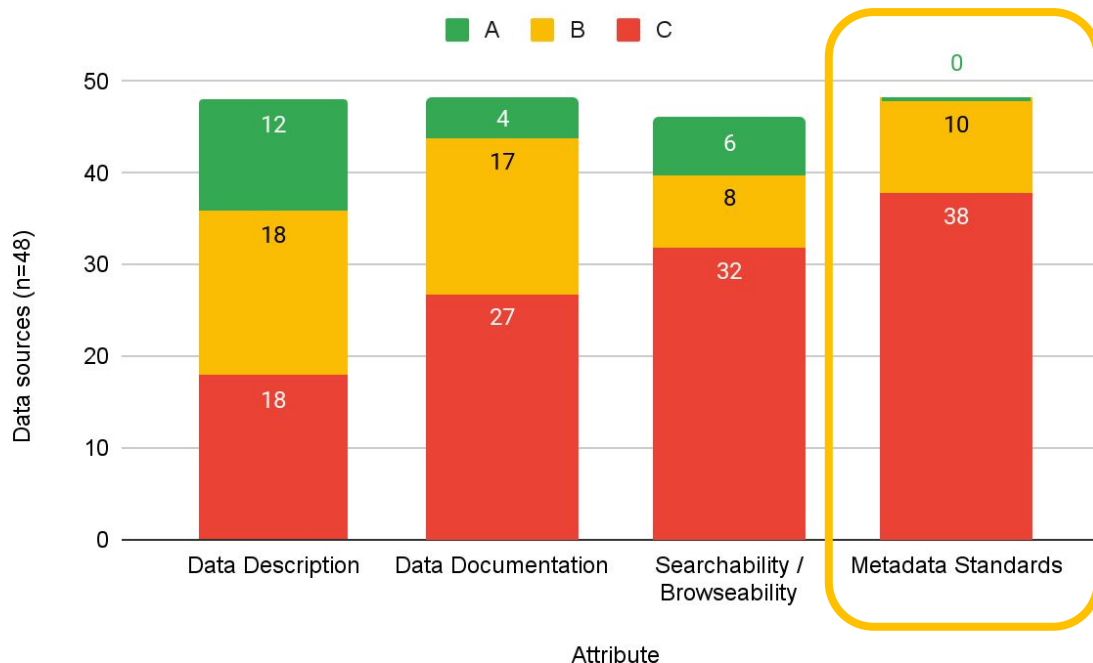
1. **Find** (as many) examples of Canadian restricted data sources
2. **Evaluate** Canadian restricted data sources based on how well they make their data discoverable and accessible
3. **Extract** metadata commonalities from restricted data sources to test alignment with existing metadata schemas

Evaluating restricted health data sources



- **79% (n=38)** received a “C” grade for metadata standards
- **0% (n=0)** received an “A” grade for metadata standards

Figure 1. Data Discovery Grades by Attribute in Restricted Health Data Sources

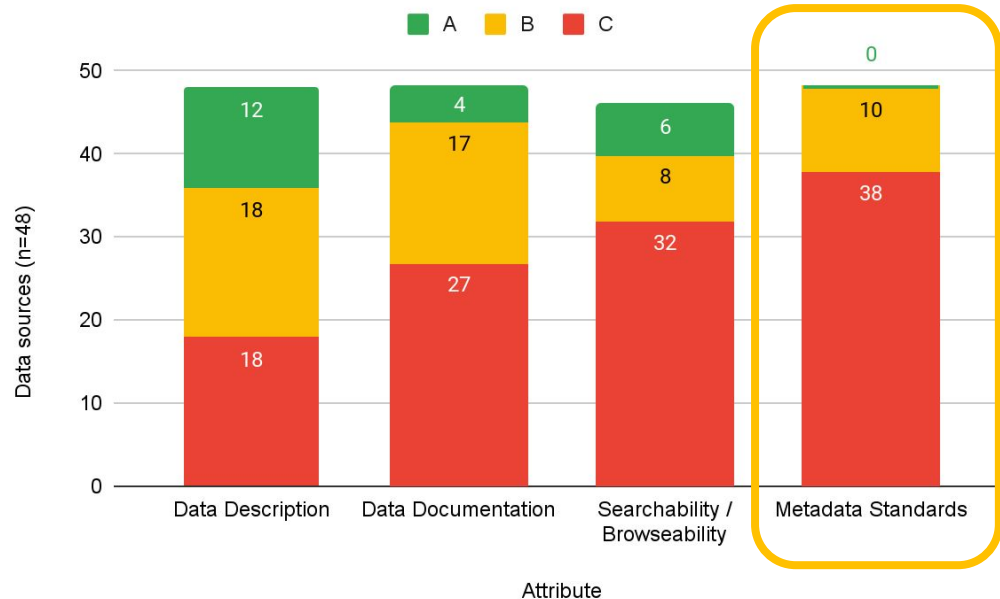


Grading exercise: Emerging questions

Do data sources describe their dataset(s) and access procedures in similar ways?



Do these similarities map to existing metadata schemas?



Project goals

1. **Find** (as many) examples of Canadian restricted data sources
2. **Evaluate** Canadian restricted data sources based on how well they make their data discoverable and accessible
3. **Extract** metadata commonalities from restricted data sources to test alignment with existing metadata schemas

Metadata “element” extraction

Review restricted health data sources to identify:

A) Dataset “elements”

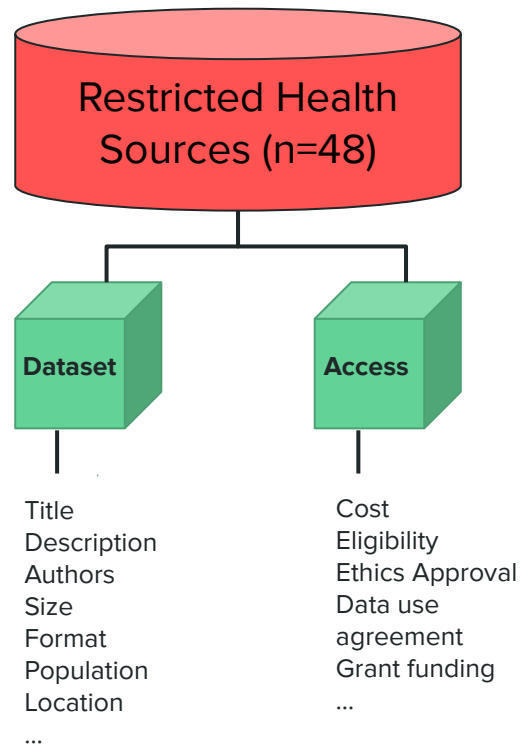
- a) Descriptive information they provide about their dataset(s) (e.g., format)

B) Access “elements”

- a) Descriptive information they provide about their data request process (e.g., application req’s)

Are there commonalities across sources?

Can existing metadata standards accommodate these commonalities?



Dataset “Elements”

Identified **35** common dataset “elements” across sources

Example: Common data characteristics used captured across sources

Metadata Entity	Sub Entity	Definition	Frequency
Data characteristics			91
	Population	Population studied	26
	Data quality	Data quality measures applied to dataset	19
	Geographic	Geographic location where data was collected	12
	Variables	Variable level information about dataset	11
	Temporal, Date range	The beginning and end date of data collection	9
	Temporal, Start date	The start date for data collection	4
	Temporal, End date	The end date for data collection	4
	Temporal, Reference period	When the data was made available for access	3
	Size of dataset	Size of dataset	3

Access “Elements”

Identified **27** common access “elements” across sources

Example: Request requirement commonalities across sources

Metadata Category	Metadata Subcategory	Access Element	Definition	Frequency
Request requirements				1765
	Research team information			668
		Requestor Name	data request	212
		PI	project for which data is being	185
		Team members	on study where data would be	96
		Primary contact	of the primary person	82
		Student info	of any students engaging with	50
		background	professional qualifications of	25
		Conflict of interest	with any of the data requestors	18
	Research plan			374
		Study purpose	purpose	113
		Study design	of study	82
		External linkages	additional data linkages that	50
		Timeline	length of study where data will	49
		Project title	project where data will be used	47
		Support needed	management, storage, security,	20
		Scientific review	of scientific review their study	13
	Request data description			184
	Ethics approval			151
		Ethics review	ethics approval has been	100
		Risks/benefits	and benefits of using data	34
		Participant recruitment	they will recruit participants for	17
	Data management			215
		Storage and security	requested data will be securely	134
		Processing	will be processed	42
		Access restrictions	restrictions are placed on	39
	Funding			96
	Intended use			48

Dataset “Element” metadata mapping

Common Metadata Alignment (35 elements)	Metadata Schemas				
	DataCite	DDI Lifecycle	DDI Codebook	DCAT	DATS
Exact	37.1% (n=13)	91.4% (n=32)	85.7% (n=30)	60% (n=21)	40% (n=14)
Partial	28.6% (n=10)	8.6% (n=3)	14.3% (n=5)	31.4% (n=11)	45.7% (n=16)
None	34.3% (n=12)	0% (n=0)	0% (n=0)	20% (n=7)	14.3% (n=5)

Dataset “Element” metadata mapping

Common Metadata Alignment (35 elements)	Metadata Schemas				
	DataCite	DDI Lifecycle	DDI Codebook	DCAT	DATS
Exact	37.1% (n=13)	91.4% (n=32)	85.7% (n=30)	60% (n=21)	40% (n=14)
Partial	28.6% (n=10)	8.6% (n=3)	14.3% (n=5)	31.4% (n=11)	45.7% (n=16)
None	34.3% (n=12)	0% (n=0)	0% (n=0)	20% (n=7)	14.3% (n=5)

Access “Element” mapping

No existing metadata schema was sufficient to describe the 27 elements we found

Schemas provide general metadata fields about access e.g., “Access Restrictions”

General fields cause two main issues:

- Data creator is not prompted to provide sufficient detail
- Data requestor does not receive clear instructions about access process

What does this all mean?

Study findings

Many Canadian restricted data sources in Canada are available for use, but are difficult to find and access

Lack of discoverability = they do not utilize metadata

But!

- Data sources describe their data and access procedures in similar ways
- Descriptions of restricted data could be accommodated by existing metadata schemas

However...

- Metadata schemas do not accommodate information about the access request process

Recommendations

Digital Research Alliance of Canada:

- Work with Canadian restricted data sources to adopt metadata
- Make datasets discoverable in aggregators, indexes, and catalogues (e.g., [Lunaris](#))

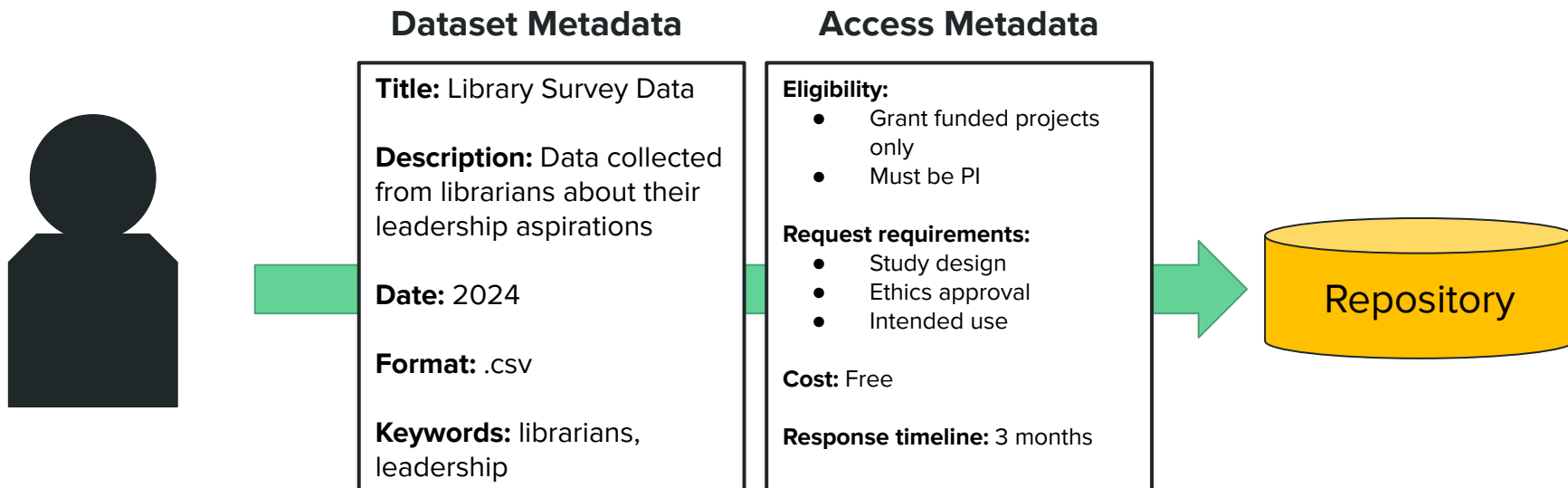
Metadata standards bodies:

- Revise metadata schemas to account for access request processes

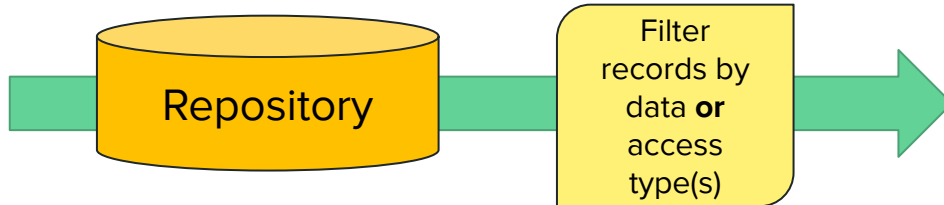
Data repositories:

- Incorporate better access metadata into systems that store restricted datasets

Imagine...researcher deposits restricted dataset



Imagine...researcher looking for data



Title: Library Survey Data

Description: Data collected from librarians about their leadership aspirations

Date: 2024

Format: .csv

Keywords: librarians, leadership

Eligibility:

- Grant funded projects only
- Must be PI

Request requirements:

- Study design
- Ethics approval
- Intended use

Cost: Free

Response timeline: 3 months

Dataset Record

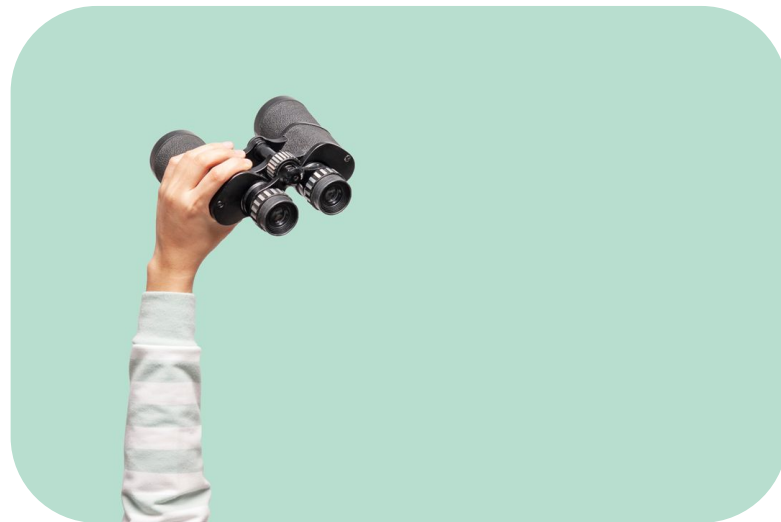
Early next steps towards discovery and access...

Incoming Tri-Agency Data Deposit Policy:

- Develop minimal metadata disclosure for researchers collecting restricted data with Tri-Agency funding
- Include both dataset *and* access metadata

Digital Research Alliance of Canada CAM Project:

- Initiative to store and make available controlled access data (i.e., restricted datasets)
- Incorporate access-specific metadata into the [Federated Research Data Repository](#) for restricted datasets



Concluding thoughts

Restricted data deserves to be made discoverable and accessible

Existing metadata schemas, technical infrastructure, and data sharing policies cannot adequately support restricted data

This work is hard – privacy, stewardship, good governance, security are crucial

More investment needed to ensure that this valuable data does not remain hidden and inaccessible

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):659–681.
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.

Questions/Comments?

kevin.read@usask.ca