Surveying the landscape of CIHR-funded research data sharing practices: An analysis of the published literature

Authors:

Corresponding Author:
Kevin B. Read, MLIS, MAS
University of Saskatchewan
kevin.read@usask.ca

Co-authors
Heather Ganshorn, MLIS
University of Calgary

Sarah Rutley, MLIS, MA
University of Saskatchewan

David R. Scott, MLIS, MA
University of Lethbridge

Funding:
N/A

Conflict of Interest:
None to declare.

Introduction

To improve health outcomes and research reproducibility, health sciences research has become increasingly focused on the production, management, and sharing of research data. The call to make health sciences research more reproducible and reusable has spearheaded a number of initiatives in the United States (1,2), Europe (3,4), and Canada (5) to improve data discoverability, accessibility, and transparency. The importance of data sharing in the health sciences has been well documented. Sharing research data improves the findability and availability of research outputs, which can spearhead new research discoveries (6–11); encourages transparency and holds the research community accountable (12–14); and improves the interoperability of data across research communities and systems (15–17).

Canada is at a crucial stage of development with respect to improving its data management and sharing initiatives. The Canadian Tri-Agency is drafting research data management (RDM) and sharing funding requirements (18), Canadian publishers have begun to release data sharing policies (19), the Federated Research Data Repository (20) has made it possible to discover data that are produced and stored in Canadian repositories, and a New Digital Research Infrastructure Organization was established to respond to emerging data needs within the Canadian digital research landscape (21). Although these efforts aim to make datasets more discoverable, valuable data shared alongside publications, in external discipline-specific repositories, via websites, or by request, are difficult to locate, access, and reuse. The availability of Canadian health sciences research data is a topic that has yet to be explored in the literature but is vital for understanding researchers' data sharing practices in a Canadian context.

As the Tri-Agency prepares to release a policy that encourages RDM and data sharing, and new initiatives are established to locate Canadian research products online, we see value in identifying how and where Canadian research data are being shared, and what steps have been taken to make these data reusable. To that end, this study aims to understand the Canadian data sharing landscape by reviewing how and where Canadian Institutes of Health Research (CIHR) funded data is shared, and comparing CIHR-funded researchers' current data sharing practices to the Tri-Agency principles for RDM and sharing (22).

Methods

Identification of CIHR-funded publications

This study identified all CIHR-funded publications within the PubMed and PubMed Central (PMC) databases that indicated they shared research data underlying their published results. Both PubMed and PMC have developed dataset search filters (23) that identify publications that indicate data underlying the results have been shared. Within the context of this study, we define research data as "data that are used as primary sources to support technical or scientific

enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results." (24)

Using PMC, this study first identified all CIHR-funded publications that included a data availability statement. Data availability statements contain the authors' description of where and how to gain access to the research data underlying the published manuscript. Additional publications were identified using PubMed's data filter, which indicates when data have been shared in a data repository. These filters were combined with CIHR-related keywords in English and French, using the grants information field from both databases (Table 1). The date range of our search strategy identified publications on or before December 31, 2019.

Table 1. Search strategy per database

Database	Search Filter	CIHR-strategy	Results
PubMed Central	"has associated data"[filter] OR "has data citations"[filter]	("canadian institutes of health research"[Grant Number] OR cihr[grant number] OR IRSC[grant number] OR "Instituts de recherche en sante du Canada"[Grant Number] OR IRSC[Grant Number])	2536
PubMed	data[filter]		2624

Metadata extraction

After removing duplicates based on overlap between PubMed and PMC, 4,988 publications remained (PMC=4039, PubMed=949). Using this sample, select metadata fields were extracted from each publication for analysis using the Open Access Subset API (25), which allows the full text metadata from a publication to be extracted under a Creative Commons license. When full text metadata was not available via the Open Access Subset, it was extracted using the minimal level of metadata available in PMC. Publications that were not available in PMC (n=949) had a limited set of metadata extracted from PubMed (Table 2). The metadata extraction process was successful in retrieving the metadata for 4,144 publications, which served as the sample for our analysis. The Python scripts used to extract the metadata are available via our Open Science Framework (OSF) Project (26).

Table 2. Extracted metadata fields from PMC, PubMed, and the PMC Open Access Subset

Metadata Field	Description	PMC metadata	PMC Open Access Subset	MEDLINE / PubMed Dataset metadata
Author affiliation	Includes the institutional affiliation and address (including email address, when available) of the authors of the publication as it appears in the journal.	x	x	x
Publication date	The date that the publication was published.	x	x	x
Journal title	The journal title abbreviation, full journal title, or ISSN number	x	x	x
Publication type	The type of publication as categorized by MEDLINE	x	x	x
Corresponding author	The name of the corresponding author of the publication	x	x	x
Data availability statements	publications or manuscripts with data availability statements		x	
Data citations	publications or manuscripts with data citations		x	
MeSH Major Topic Headings	A MeSH term that is one of the main topics discussed in the publication.	x	x	x
publication body - Key Terms	Includes all key terms in the body of a publication except for the Abstract and References.		x	

Acknowledgements	Includes all words in the acknowledgement section of a publication (e.g., “National Institutes of Health[ack]”).	x	x	
Grant number	The grant number search field includes research grant numbers, contract numbers, or both that designate financial support by Agency of the US PHS (Public Health Service), and other national or international funding sources.	x	x	x

Examination of CIHR-funded data sharing practices

Using the extracted metadata, we analyzed each publication (n=4,144) using descriptive statistics to explore data accessibility; how, where and by whom research data was shared; and the inclusion of documentation to support data reuse.

Our data collection instrument and descriptive statistics were generated and captured in a REDCap database. The instrument and data dictionary used for our final analysis are available in our OSF project (26).

Data sharing status

To frame our analysis, we grouped CIHR-funded data sharing practices into four categories representing the most commonly identified data sharing status types (Table 3). We examined the frequency of each category across our entire sample (n=4,144) and over time.

Table 3. Data sharing status categories and their definitions

Status Category	Definition
(1) Data accessible	Research data files could be identified, accessed, and downloaded.
(2) Data available	Authors stated either within the manuscript, the data availability statement, or the acknowledgements that research data was available upon request or via an application process.
(3) Data sharing not applicable/possible	Authors stated either within the manuscript, data availability statement, or acknowledgements that research data sharing was not possible or applicable.
(4) No evidence of data	Authors made no mention of data sharing within the

sharing	manuscript, data availability statement, or acknowledgements; indicated data sharing would be available at a future date; or the publication contained no research data files.

Data sharing methods

Using the metadata available (Table 2) and building upon the high level data sharing status categories identified (Table 3), we recorded all methods of data sharing evident within each publication. Methods of data sharing included but were not limited to sharing data via a data repository, within the supplementary files, by request or application, within the publication, via a website, or when an author stated data sharing was not applicable or possible.

If an author's data sharing statement indicated that an application was required to access the data, we captured all reasons why authors insisted on this requirement. Similarly, if an author stated that data could not be shared at all, we captured all reasons provided why this was the case.

Finally, we examined whether data sharing statements made by authors within a publication aligned with how data were shared in practice. When authors stated that all research data needed to understand the results were within the publication, we reviewed the publication for evidence that no additional research data files were needed to understand the findings. When authors stated that research data were available in the supplementary files of a publication, we attempted to locate and access the data within the supplementary files section. We documented instances of misalignment between author statements and if and/or how data were shared, as well as when we were unclear about whether author statements reflected data sharing accurately.

Research data documentation

To expand our analysis, we explored the types of documentation that were included alongside accessible and available research data (Table 3, Categories 1 and 2). We identified types of documentation based on the Tri-Agency Statement of Principles on Digital Data Management (22), which makes recommendations on adherence to standards, data collection and storage, and metadata documentation. We then examined each publication to determine whether or not documentation such as study protocols, data analysis plans, software and/or code, data dictionaries, readme files, data collection instruments, videos, or data management plans was provided. Documentation of this kind has been identified as necessary for improving the transparency, reproducibility, and reusability of research results (27–30). Recording the presence of these files also enabled an analysis of the frequency of documentation inclusion over time.

The CIHR-funded data sharing landscape

Our study identified the institutions that most frequently share research data and the journals where CIHR-funded data sharing frequently occurs. Institutions and journals were categorized and ranked according to their data sharing status (Table 3).

All data collected during the present study were exported from the REDCap database and analyzed using Stata/SE 16.0 software. The raw data extracted from PubMed and PMC, the synthesized data exported from REDCap, and the analyzed data from Stata along with a summary analysis report are available in our OSF Project (26).

Results

Of the 4,144 CIHR-funded publications included in this study, 45% made their data accessible, 22% made their data available (via request or application), 7% indicated data sharing was not applicable or possible, and despite isolating our sample to publications that had indicated data sharing of some kind, 38% provided no evidence of data sharing (Table 4). Note that these categories are not mutually exclusive, as many publications shared multiple datasets in different ways. Figure 1 demonstrates the extent of overlap between the four data sharing status types, and Figure 2 examines the frequency of these four categories over time.

Table 4. Frequency of data sharing status -- not mutually exclusive to a single publication (n=4,144)

Data sharing method	Frequency (n)	Percent (%)
Data accessible	1,876	45.27
No evidence of data sharing	1,558	37.60
Data available	909	21.94
Data sharing not applicable or possible	304	7.34

Figure 1. Frequency of publications categorized by data sharing status (n=4,144)

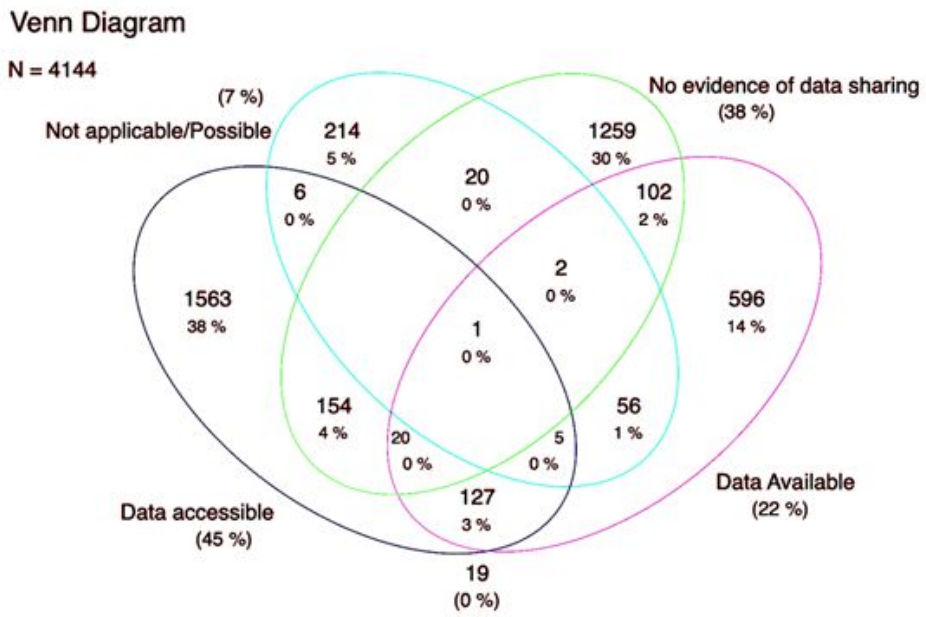
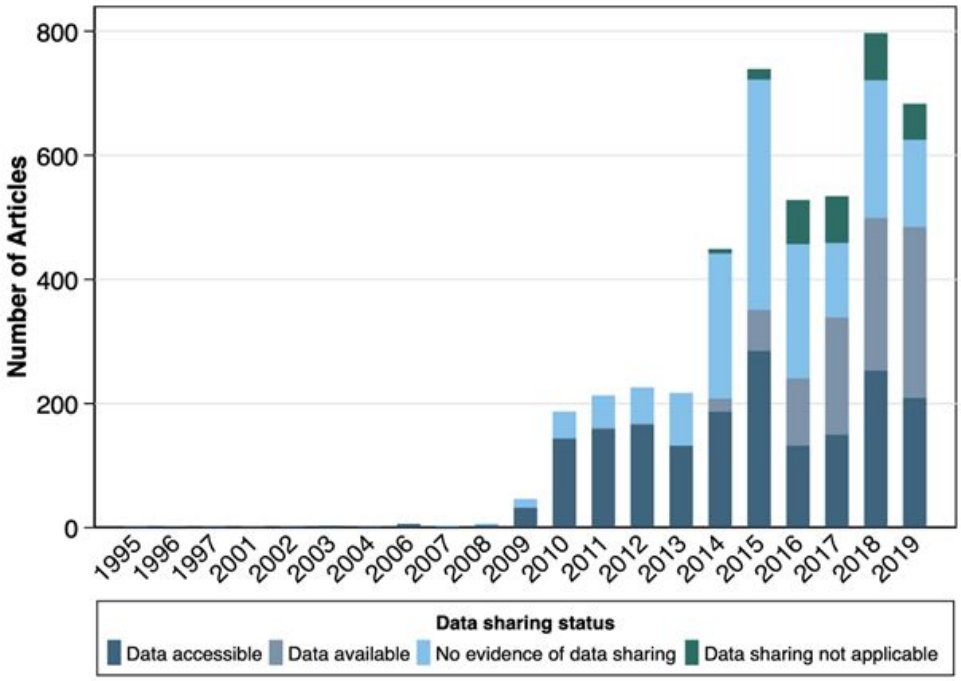


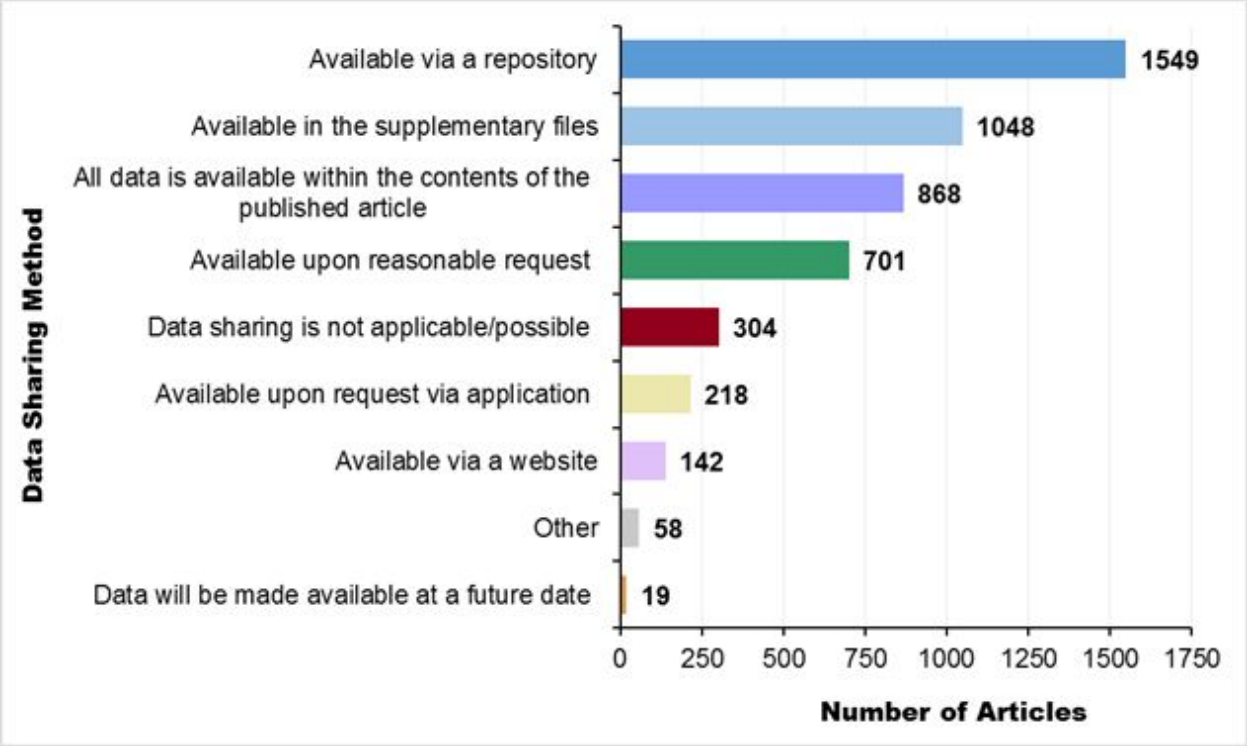
Figure 2. Frequency of data sharing status category over time -- not mutually exclusive to a single publication (n=4,144)



Data sharing methods

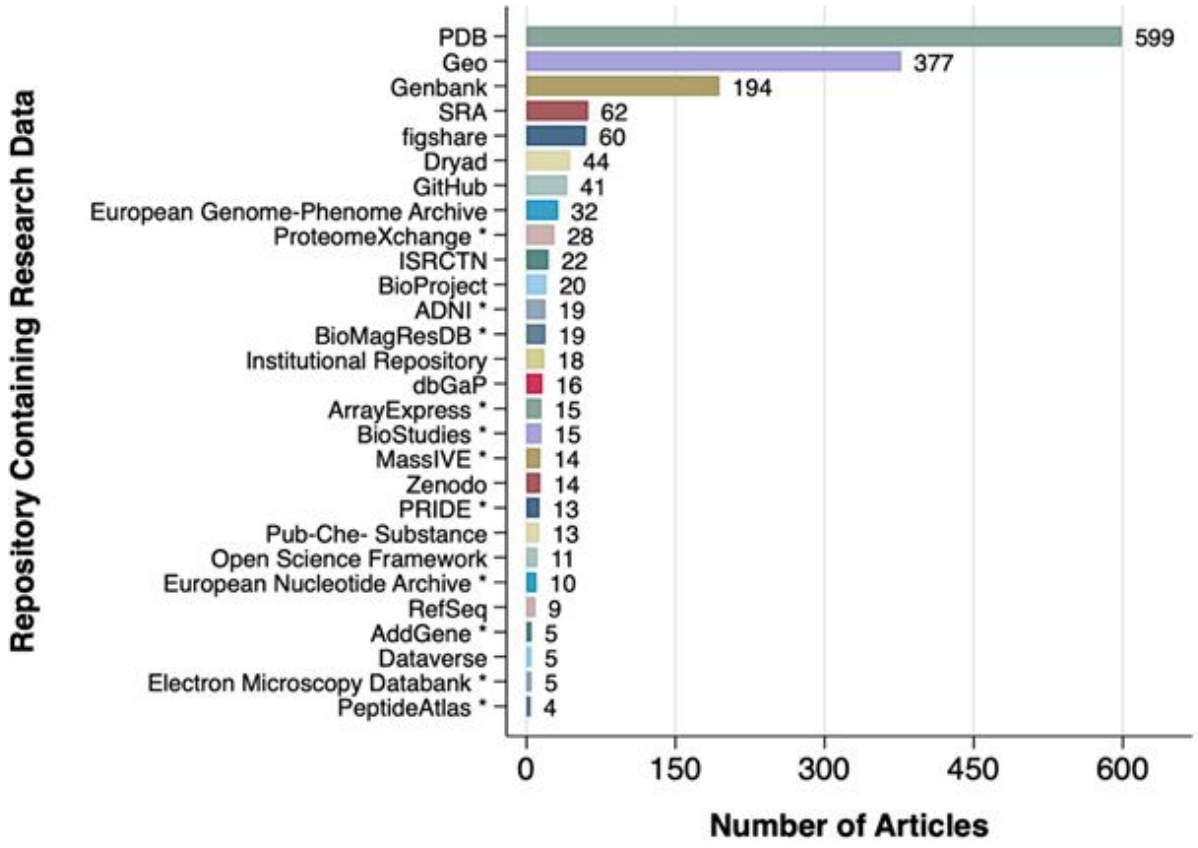
The most frequent method of data sharing was via a repository (37%) followed by within the supplementary files (25%). Notably, 22% of articles stated that data were available either by request (17%) or application (5%), despite providing little detail about how to acquire these data. 21% of publications stated that all data underlying results were available within the content of the publication. 13% of publications had no evidence of or information about data sharing whatsoever. Some publications shared data in multiple formats, and therefore may be represented in more than one category. The types and frequency of all data sharing methods are illustrated in Figure 3.

Figure 3. Frequency of data sharing method -- not mutually exclusive to a single publication (n=4,144)



Among publications that reported data sharing via a repository (n=1,549), there were 97 distinct repositories represented (see analysis report on the OSF for complete listing) (26). The most represented repositories were the Protein Data Bank (PDB) (n=599, 39%), Gene Expression Omnibus (GEO) (n=377, 24%), and GenBank (n=194, 13%). A complete breakdown of repositories is shown in Figure 4.

Figure 4. Frequency of data repositories used to store CIHR-funded research data (n=1,544).

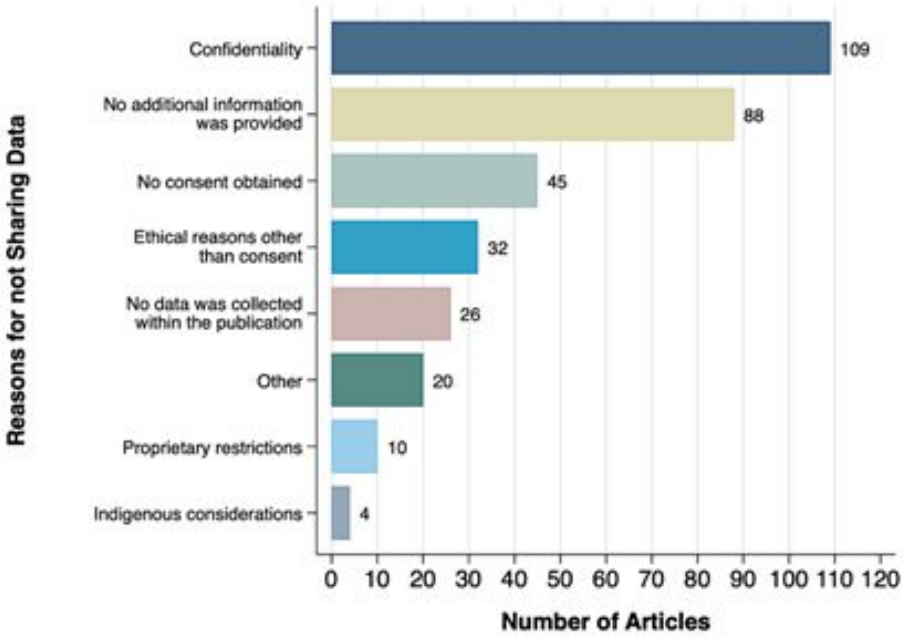


**Asterisks denote repositories that were not anticipated within our instrument, and were added from a free-text "Other" category during data collection.*

234 publications indicated that an application was required to access the data underlying the results. The most frequent justification for this requirement was the need to complete a data access, transfer or use agreement (28%), followed by general ethics concerns (24%), confidentiality (21%), license restrictions (10%), and Indigenous considerations (2.5%). Nearly 10% of publications provided no explanation for why an application was required. Among publications that required an application, none included metadata sufficiently outlining the requirements for access and approval.

Among publications that indicated that data sharing was not applicable or possible (n=300), the most common reason cited for being unable to share data was confidentiality (36%). Over 29% of publications that indicated data sharing was not applicable or possible provided no justification at all. A complete list of reasons why data could not be shared is available in Figure 5.

Figure 5. Frequency of reasons for not sharing data (n=300)



Finally, our comparison of author data availability statements with actual data sharing practices revealed that 71.8% (n=752) of publications reporting data available in the supplementary files (n=1,048) did not share data in this way. Similarly, 39.7% (n=345) of statements that all data were available within the publication (n=870) were flagged on the grounds that although there was clear evidence of data collection, no data was shared within the publication or supplementary files. The authors of this study agreed that in many cases authors may have incorrectly considered tables and figures to be research data.

Research data documentation

The documentation provided alongside the publications in our sample was varied, with supplementary figures and/or tables, study protocols, research data files, and transparent reporting forms most frequently represented (Table 5).

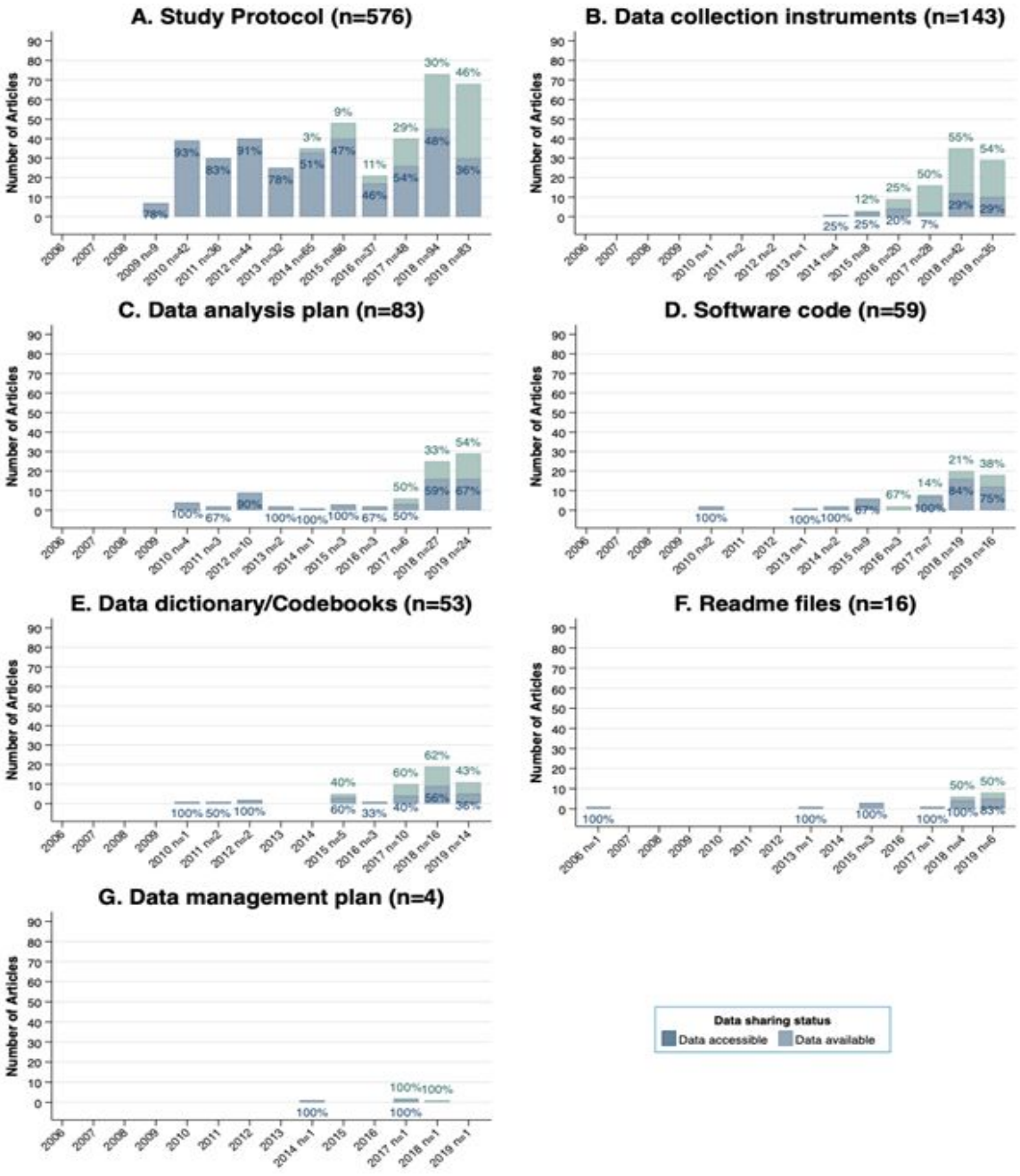
Table 5. Documentation identified and categorized by data sharing status

Type of Information	Data sharing status (by count)			
	Data Accessible	Data Available	No evidence of data sharing	Data sharing not possible/applicable
Supplementary figures and/or tables	1,258	385	881	65

Study protocol	332	94	200	18
Data files	504	82	3	10
Transparent reporting form	138	57	107	4
Data collection instruments	31	62	52	39
Videos	56	20	41	1
Data analysis plan/documentation	58	25	12	4
Image files	23	5	43	0
Software code	46	13	8	2
Data dictionary/codebook	26	24	10	6
Preservation formats for structured data	50	4	2	0
Readme files	15	5	1	1
Data management plan	2	2	0	0
Other	29	30	28	6

Referring to the recommended documentation types outlined in the Tri-Agency Data Management Principles (22), we examined how frequently these types of documentation were included alongside publications that made data accessible or indicated that data were available (Table 3, Categories 1 and 2), over time (Figure 6). Our findings indicate that the types of documentation required to understand and reuse research data are seldom provided in CIHR-funded publications that share data (13%, n=554). Study protocols were the most frequently included at 13.9% and data management plans were least frequent at 0.1%. Although documentation supporting reuse was scarce, the practice of including data-related documentation has grown in the past three years, with our results showing the increasing availability of data analysis plans, software code, and data collection instruments.

Figure 6. Frequency of data management-specific documentation grouped by data sharing status over time (n=4,144)

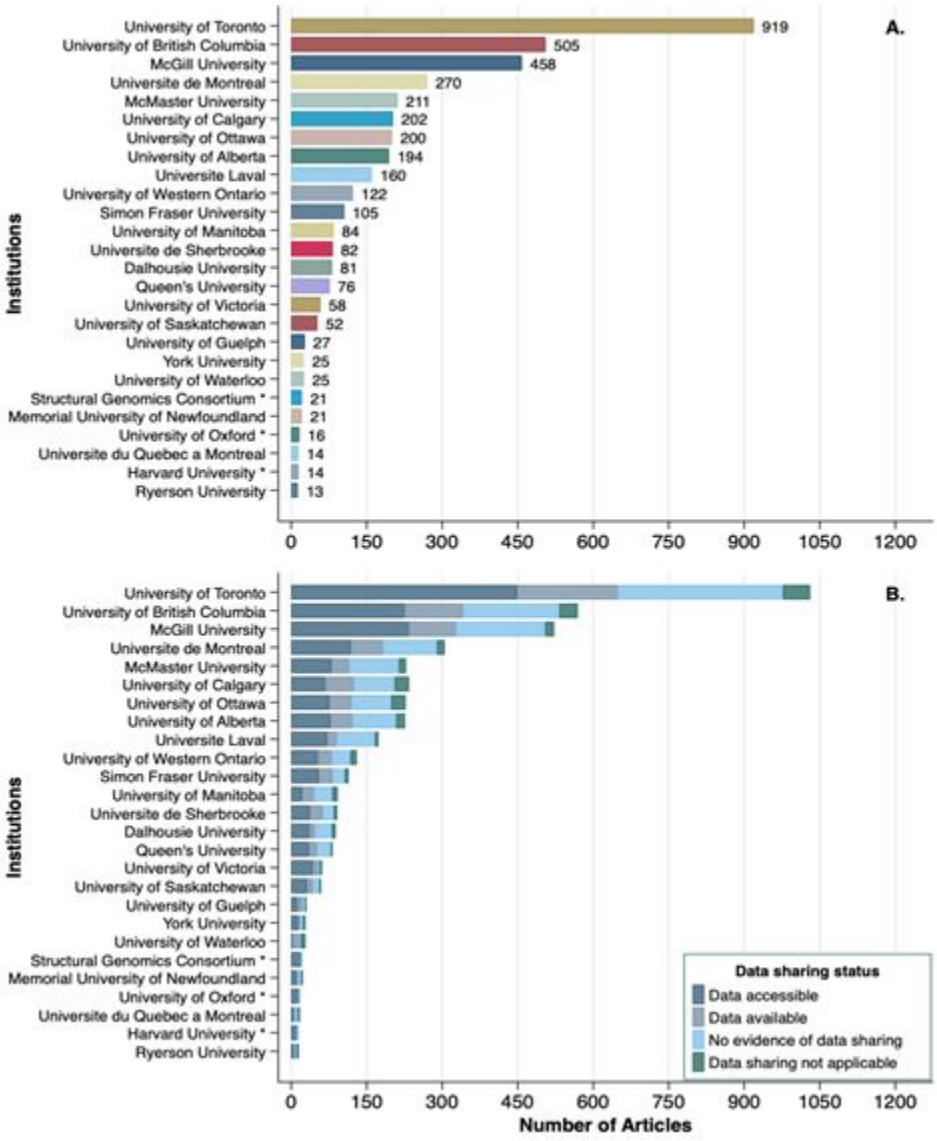


The CIHR-funded data sharing landscape

Institutions associated with the CIHR-funded publications included in our sample are shown in Figure 7A and 7B. The University of Toronto had the greatest proportion of CIHR-funded publications (22.18%, n=919). Among institutions with more than 10 publications, those with the

greatest proportion of publications where data were accessible (Table 3, Category 1) or available (Table 3, Category 2) were the Structural Genomics Consortium (95.2%, n=20) and the University of Waterloo (48%, n=12), respectively.

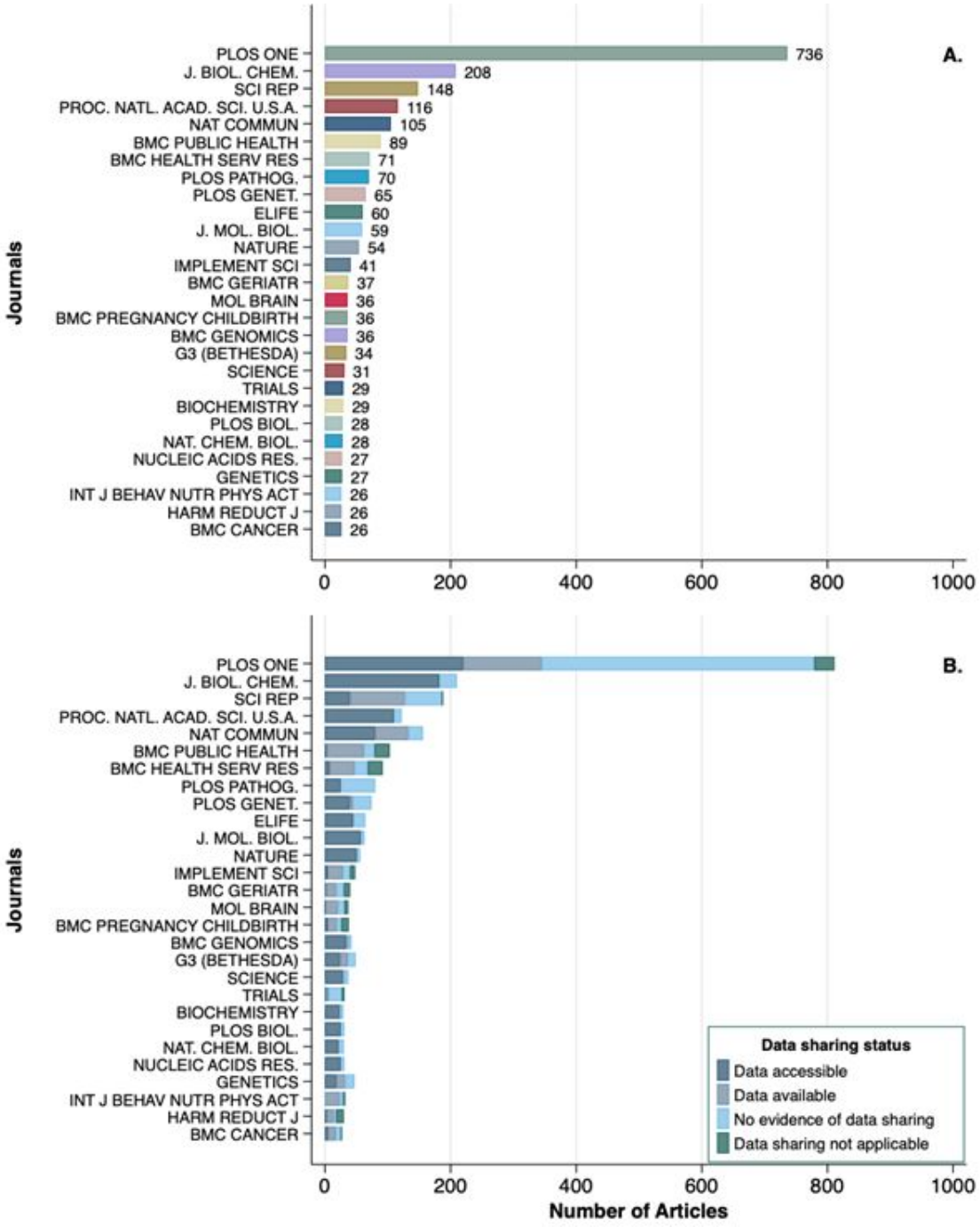
Figure 7A. Frequency of publications identified by institution (n=4,144)
Figure 7B. Frequency of publications identified by institution grouped by data sharing status (n=4,144)



Of the 4,144 publications, the journals used most frequently were PLOS One (17.76%, n=736) followed by the Journal of Biological Chemistry (5.02%, n=208). Among journals with more than 25 CIHR-funded published publications, the top three journals that included examples of accessible data (Table 3, Category 1) were the Journal of Molecular Biology, (96.6%), Molecular Cell (95.7%), and Proceedings of the National Academy of Sciences, U.S.A. (94.8%). Where data were available (Table 3, Category 2), the International Journal of Behavioral Nutrition and

Physical Activity (88.46%) and BMC Medical Research Methodology (73.9%) were the most prominent (Figure 8A and 8B).

Figure 8A. Frequency of publications identified by journal (n=4,144)
Figure 8B. Frequency of publications identified categorized by data sharing status (n=4,144)



Interpretation

This study highlights significant room for growth in improving the discoverability, accessibility, and usability of CIHR-funded research data. While, encouragingly, repositories were the most common venues authors chose to share their data (37%), the remainder of shared data was made available within the publication or its supplementary files, by request or application, via a website, or by a long tail of other means (see analysis report on OSF) (26). Sharing data in a repository is recognized as best practice because it provides infrastructure for discovery, structured metadata, and long-term, reliable access. The other sharing methods we identified present by comparison significant barriers to discovery, interpretability, and access, in that they make data difficult to find, do not incorporate metadata, and do not facilitate data access consistently. These characteristics conflict with expectations outlined in Canada's Tri-Agency data management (22) and international FAIR guiding principles (15).

Metadata is an essential component of data sharing that provides valuable context about the nature of data, how they were collected, and how they can be reused (31). Metadata also improve discoverability by applying structured descriptors to data that allow them to be searched for and retrieved. Most sharing methods we encountered during our study did not incorporate metadata, making data difficult to locate, interpret, and reuse. Inadequate metadata descriptions are a recognized problem in the data sharing landscape (32–34), and the CIHR-funded data sharing practices assessed in this study are no different. Without adequate metadata, these data will remain hidden within the publication and their utility for future research remains in question (35).

In instances where authors indicated that data was available by request or application (22%), they did not provide adequate instructions on how to formally acquire the data, leaving interested researchers with no guidance on what a successful request would look like. The absence of metadata elaborating on application requirements calls into question the true availability of these data, and impedes future research based upon them. The challenges of requesting access to data in the health sciences have been studied elsewhere in relation to the inadequate transparency and standardization of data use agreements (33,36,37). Because data available by request or application are often collected from human participants, improving the discoverability of and access to these sensitive data will help prevent unnecessary study replication, create opportunities for pooling related data, and increase research efficiency to accelerate new discoveries (33,34,36,37). Our findings indicate that current practices in CIHR-funded data sharing lack the standardization and transparency necessary to secure these outcomes.

In our examination of practices that support reusability, we found that the most frequent types of documentation shared alongside data – supplementary figures and/or tables and study protocols – do not generally support the interpretation and reuse of data. Tables and figures were often reiterations of visualizations presented within the body of the publication; study protocols and data collection instruments, while helpful for contextualizing how data was gathered and analyzed, do little to help others understand the data themselves and how to

interact with them. Descriptive documentation for data such as codebooks and data dictionaries, and actionable supporting files such as code and software, are increasingly recognized as necessary components of human- and machine-readable research data (27,30,38), and our results indicate that CIHR-funded research data sharing practices can vastly improve in this area.

As Canada implements new data sharing requirements for federally funded research, our study highlights the importance of developing policies and standards at the federal and institutional levels to ensure that all research data underlying published findings have quality metadata attached to them, and that sufficient documentation to support interpretation and reuse is provided. Future directions of study should focus on the development of metadata standards for sensitive data to facilitate reuse and support transparent data request processes. We also recommend that Canadian data repositories explore how to better accommodate sensitive data so that they can be made discoverable while honouring access restrictions and privacy requirements.

Limitations of the study

This study used a sample of CIHR-funded publications pulled exclusively from PubMed and PMC. While these are the most comprehensive biomedical databases available, there are likely other databases where CIHR-funded publications exist. To manage study feasibility, we limited our review of documentation to that which was shared or stated within the publication and did not extend this analysis to repositories or websites where some research data was shared.

Conclusion

This study surveys the complex landscape of CIHR-funded data sharing practices, revealing a diverse range of data sharing methods and, in 38% of cases, an absence of data sharing altogether. It is remarkable that over 70% of publications that shared data did not incorporate sufficient metadata or documentation to facilitate discovery, access, and reuse. Without policies and standards in place that anticipate the upcoming Tri-Agency data management policy, and enhanced support for researchers seeking to implement best practices in data management and sharing, the majority of publicly funded research data will remain hidden, inaccessible, and unusable. For CIHR-funded data in particular, transparent metadata and reporting guidelines for sensitive data will be essential for improving data discoverability and accessibility across the health sciences.

Data Availability Statement

All raw, processed, and analyzed data, as well as accompanying documentation, reports and scripts are available on the Open Science Framework at https://osf.io/n9jv5.

Acknowledgements

We would like to thank the Canadian Hub for Applied and Social Research (CHASR) for their support in procuring and analyzing the research data in this study. This project was supported in part by the University of Saskatchewan Faculty Recruitment and Retention Program.

References

1. Collins, Francis. Statement on Final NIH Policy for Data Management and Sharing. National Institutes of Health. 2020.

2. FS Collins, Tabak L. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612–3.

3. Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, et al. Identifying ELIXIR Core Data Resources. F1000Research. 2016 Sep 30;5:ELIXIR-2422.

4. Artini M, Atzori C, Bardi A, Bruzzo SL, Manghi P, Mannocci A. The OpenAIRE Literature Broker Service for Institutional Repositories. -Lib Mag. 2015 Dec 11;21(11/12):95–104.

5. OpenAIRE. OpenAIRE joins forces with Canada's federal granting agencies and CARL [Internet]. Vol. 2019, OpenAIRE. OpenAIRE; 2019. Available from: https://www.openaire.eu/openaire-joins-forces-with-canada-s-federal-granting-agencies-and-carl

6. Humphreys GS, Tinto H, Barnes KI. Strength in Numbers: The WWARN Case Study of Purpose-Driven Data Sharing. Am J Trop Med Hyg. 2019;100(1):13–5.

7. McAlister VC, Harvey EJ. The benefits and risks of requiring researchers to share data. Can J Surg. 2016;59(6):364–5.

8. Owens B. Data sharing: Access all areas. Nature. 2016;533(7602):S71-2.

9. Mendelson DS, Bak PRG, Menschik E, Siegel E. Informatics in radiology: image exchange: IHE and the evolution of image sharing. Radiogr Rev Publ Radiol Soc N Am Inc. 2008 Dec;28(7):1817–33.

10. Steele Gray C, Barnsley J, Gagnon D, Belzile L, Kenealy T, Shaw J, et al. Using information communication technology in models of integrated community-based primary health care: learning from the iCOACH case studies. Implement Sci IS. 2018 26;13(1):87.

11. Vestrup JA, Phang PT, Vertesi L, Wing PC, Hamilton NE. The utility of a multicenter regional trauma registry. J Trauma. 1994 Sep;37(3):375–8.

12. Dumontier M, Wesley K. Advancing Discovery Science with FAIR Data Stewardship: Findable, Accessible, Interoperable, Reusable. Ser Libr. 2018;74(1–4):39–48.

13. Reiser L, Harper L, Freeling M, Han B, Luan S. FAIR: a call to make published data more findable, accessible, interoperable, and reusable. Mol Plant. 2018;11(9):1105–8.

14. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251).

15. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3.

16. Persaud N. A national electronic health record for primary care. CMAJ Can Med Assoc J J Assoc Medicale Can. 2019 Jan 14;191(2):E28–9.

17. Rathi VK, Strait KM, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. Trials. 2014 Oct 2;15:384.

18. Government of Canada. Draft Tri-Agency Research Data Management Policy for Consultation [Internet]. Government of Canada Policies and Guidelines. 2018. Available from: http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html

19. Kelsall D. New CMAJ policy on sharing study data. CMAJ Can Med Assoc J. 2017;189(34):E1082–E1082.

20. Wilson L. Exploring the Canadian Federated Research Data Repository Service. Biodivers Inf Sci Stand. 2017;

21. Government of Canada. Digital Research Infrastructure [Internet]. Government of Canada Innovation, Science, and Economic Development Canada. 2019. Available from: http://www.ic.gc.ca/eic/site/136.nsf/eng/home

22. Government of Canada. Tri-Agency Statement of Principles on Digital Data Management [Internet]. Government of Canada Policies and Guidelines. 2016. Available from: http://science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html

23. National Library of Medicine. Data Filters in PMC and PubMed [Internet]. NLM Technical Bulletin. 2018. Available from: https://www.nlm.nih.gov/pubs/techbull/ma18/brief/ma18_pmc_data_filters.html

24. CASRAI. Research data definition [Internet]. CASRAI Glossary. 2019. Available from: https://casrai.org/term/research-data/

25. National Library of Medicine. Open Access Subset [Internet]. PMC Tools. 2019. Available from: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

26. Read K, Ganshorn H, Rutley S, Scott D. Surveying the landscape of CIHR-funded research data sharing practices: An analysis of the published literature [Internet]. Open Science Framework; 2020. Available from: https://osf.io/n9jv5

27. Bakken S. The journey to transparency, reproducibility, and replicability. J Am Med Inform Assoc. 2019;26(3):185–7.

28. Holub P, Kohlmayer F, Prasser F, Mayrhofer MT, Schlünder I, Martin GM, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreservation Biobanking. 2018 Apr;16(2):97–105.

29. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain. 2020;13(1):24–24.

30. Walters WP. Code Sharing in the Open Science Era. J Chem Inf Model. 2020 Oct 26;60(10):4417–20.

31. Qin J, Ball A, Greenberg J. Functional and Architectural Requirements for Metadata. In: International Conference on Dublin Core and Metadata Applications. 2012. p. 10.

32. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform. 2020 Jul;107:103421.

33. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, et al. Barriers to accessing public cancer genomic data. Sci Data. 2019 Jun 20;6(1):98.

34. Vassar M, Jellison S, Wendelbo H, Wayant C. Data sharing practices in randomized trials of addiction interventions. Addict Behav. 2020 Mar;102:106193.

35. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data. 2020;7(1):5–8.

36. Miller J, Ross JS, Wilenzick M, Mello MM. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. BMJ. 2019 Jul 10;366:l4217.

37. Shabani M, Obasa M. Transparency and objectivity in governance of clinical trials data sharing: Current practices and approaches. Clin Trials Lond Engl. 2019 Oct;16(5):547–51.

38. Davenport JH, Grant J, Jones CM. Data Without Software Are Just Numbers. Data Sci J. 2020;19(1).

Surveying the landscape of CIHR-funded research data sharing practices: An analysis of the published literature

Authors:

Corresponding Author:
Kevin B. Read, MLIS, MAS
University of Saskatchewan
kevin.read@usask.ca

Co-authors
Heather Ganshorn, MLIS
University of Calgary

Sarah Rutley, MLIS, MA
University of Saskatchewan

David R. Scott, MLIS, MA
University of Lethbridge

Funding:
N/A

Conflict of Interest:
None to declare.

Introduction

To improve health outcomes and research reproducibility, health sciences research has become increasingly focused on the production, management, and sharing of research data. The call to make health sciences research more reproducible and reusable has spearheaded a number of initiatives in the United States (1,2), Europe (3,4), and Canada (5) to improve data discoverability, accessibility, and transparency. The importance of data sharing in the health sciences has been well documented. Sharing research data improves the findability and availability of research outputs, which can spearhead new research discoveries (6–11); encourages transparency and holds the research community accountable (12–14); and improves the interoperability of data across research communities and systems (15–17).

Canada is at a crucial stage of development with respect to improving its data management and sharing initiatives. The Canadian Tri-Agency is drafting research data management (RDM) and sharing funding requirements (18), Canadian publishers have begun to release data sharing policies (19), the Federated Research Data Repository (20) has made it possible to discover data that are produced and stored in Canadian repositories, and a New Digital Research Infrastructure Organization was established to respond to emerging data needs within the Canadian digital research landscape (21). Although these efforts aim to make datasets more discoverable, valuable data shared alongside publications, in external discipline-specific repositories, via websites, or by request, are difficult to locate, access, and reuse. The availability of Canadian health sciences research data is a topic that has yet to be explored in the literature but is vital for understanding researchers' data sharing practices in a Canadian context.

As the Tri-Agency prepares to release a policy that encourages RDM and data sharing, and new initiatives are established to locate Canadian research products online, we see value in identifying how and where Canadian research data are being shared, and what steps have been taken to make these data reusable. To that end, this study aims to understand the Canadian data sharing landscape by reviewing how and where Canadian Institutes of Health Research (CIHR) funded data is shared, and comparing CIHR-funded researchers' current data sharing practices to the Tri-Agency principles for RDM and sharing (22).

Methods

Identification of CIHR-funded publications

This study identified all CIHR-funded publications within the PubMed and PubMed Central (PMC) databases that indicated they shared research data underlying their published results. Both PubMed and PMC have developed dataset search filters (23) that identify publications that indicate data underlying the results have been shared. Within the context of this study, we define research data as "data that are used as primary sources to support technical or scientific

enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results." (24)

Using PMC, this study first identified all CIHR-funded publications that included a data availability statement. Data availability statements contain the authors' description of where and how to gain access to the research data underlying the published manuscript. Additional publications were identified using PubMed's data filter, which indicates when data have been shared in a data repository. These filters were combined with CIHR-related keywords in English and French, using the grants information field from both databases (Table 1). The date range of our search strategy identified publications on or before December 31, 2019.

Metadata extraction

After removing duplicates based on overlap between PubMed and PMC, 4,988 publications remained (PMC=4039, PubMed=949). Using this sample, select metadata fields were extracted from each publication for analysis using the Open Access Subset API (25), which allows the full text metadata from a publication to be extracted under a Creative Commons license. When full text metadata was not available via the Open Access Subset, it was extracted using the minimal level of metadata available in PMC. Publications that were not available in PMC (n=949) had a limited set of metadata extracted from PubMed (Table 2). The metadata extraction process was successful in retrieving the metadata for 4,144 publications, which served as the sample for our analysis. The Python scripts used to extract the metadata are available via our Open Science Framework (OSF) Project (26).

Examination of CIHR-funded data sharing practices

Using the extracted metadata, we analyzed each publication (n=4,144) using descriptive statistics to explore data accessibility; how, where and by whom research data was shared; and the inclusion of documentation to support data reuse.

Our data collection instrument and descriptive statistics were generated and captured in a REDCap database. The instrument and data dictionary used for our final analysis are available in our OSF project (26).

Data sharing status

To frame our analysis, we grouped CIHR-funded data sharing practices into four categories representing the most commonly identified data sharing status types (Table 3). We examined the frequency of each category across our entire sample (n=4,144) and over time.

Data sharing methods

Using the metadata available (Table 2) and building upon the high level data sharing status categories identified (Table 3), we recorded all methods of data sharing evident within each

publication. Methods of data sharing included but were not limited to sharing data via a data repository, within the supplementary files, by request or application, within the publication, via a website, or when an author stated data sharing was not applicable or possible.

If an author's data sharing statement indicated that an application was required to access the data, we captured all reasons why authors insisted on this requirement. Similarly, if an author stated that data could not be shared at all, we captured all reasons provided why this was the case.

Finally, we examined whether data sharing statements made by authors within a publication aligned with how data were shared in practice. When authors stated that all research data needed to understand the results were within the publication, we reviewed the publication for evidence that no additional research data files were needed to understand the findings. When authors stated that research data were available in the supplementary files of a publication, we attempted to locate and access the data within the supplementary files section. We documented instances of misalignment between author statements and if and/or how data were shared, as well as when we were unclear about whether author statements reflected data sharing accurately.

Research data documentation

To expand our analysis, we explored the types of documentation that were included alongside accessible and available research data (Table 3, Categories 1 and 2). We identified types of documentation based on the Tri-Agency Statement of Principles on Digital Data Management (22), which makes recommendations on adherence to standards, data collection and storage, and metadata documentation. We then examined each publication to determine whether or not documentation such as study protocols, data analysis plans, software and/or code, data dictionaries, readme files, data collection instruments, videos, or data management plans was provided. Documentation of this kind has been identified as necessary for improving the transparency, reproducibility, and reusability of research results (27–30). Recording the presence of these files also enabled an analysis of the frequency of documentation inclusion over time.

The CIHR-funded data sharing landscape

Our study identified the institutions that most frequently share research data and the journals where CIHR-funded data sharing frequently occurs. Institutions and journals were categorized and ranked according to their data sharing status (Table 3).

All data collected during the present study were exported from the REDCap database and analyzed using Stata/SE 16.0 software. The raw data extracted from PubMed and PMC, the synthesized data exported from REDCap, and the analyzed data from Stata along with a summary analysis report are available in our OSF Project (26).

Results

Of the 4,144 CIHR-funded publications included in this study, 45% made their data accessible, 22% made their data available (via request or application), 7% indicated data sharing was not applicable or possible, and despite isolating our sample to publications that had indicated data sharing of some kind, 38% provided no evidence of data sharing (Table 4). Note that these categories are not mutually exclusive, as many publications shared multiple datasets in different ways. Figure 1 demonstrates the extent of overlap between the four data sharing status types, and Figure 2 examines the frequency of these four categories over time.

Data sharing methods

The most frequent method of data sharing was via a repository (37%) followed by within the supplementary files (25%). Notably, 22% of articles stated that data were available either by request (17%) or application (5%), despite providing little detail about how to acquire these data. 21% of publications stated that all data underlying results were available within the content of the publication. 13% of publications had no evidence of or information about data sharing whatsoever. Some publications shared data in multiple formats, and therefore may be represented in more than one category. The types and frequency of all data sharing methods are illustrated in Figure 3.

Among publications that reported data sharing via a repository (n=1,549), there were 97 distinct repositories represented (see analysis report on the OSF for complete listing) (26). The most represented repositories were the Protein Data Bank (PDB) (n=599, 39%), Gene Expression Omnibus (GEO) (n=377, 24%), and GenBank (n=194, 13%). A complete breakdown of repositories is shown in Figure 4.

234 publications indicated that an application was required to access the data underlying the results. The most frequent justification for this requirement was the need to complete a data access, transfer or use agreement (28%), followed by general ethics concerns (24%), confidentiality (21%), license restrictions (10%), and Indigenous considerations (2.5%). Nearly 10% of publications provided no explanation for why an application was required. Among publications that required an application, none included metadata sufficiently outlining the requirements for access and approval.

Among publications that indicated that data sharing was not applicable or possible (n=300), the most common reason cited for being unable to share data was confidentiality (36%). Over 29% of publications that indicated data sharing was not applicable or possible provided no justification at all. A complete list of reasons why data could not be shared is available in Figure 5.

Finally, our comparison of author data availability statements with actual data sharing practices revealed that 71.8% (n=752) of publications reporting data available in the supplementary files (n=1,048) did not share data in this way. Similarly, 39.7% (n=345) of statements that all data

were available within the publication (n=870) were flagged on the grounds that although there was clear evidence of data collection, no data was shared within the publication or supplementary files. The authors of this study agreed that in many cases authors may have incorrectly considered tables and figures to be research data.

Research data documentation

The documentation provided alongside the publications in our sample was varied, with supplementary figures and/or tables, study protocols, research data files, and transparent reporting forms most frequently represented (Table 5).

Referring to the recommended documentation types outlined in the Tri-Agency Data Management Principles (22), we examined how frequently these types of documentation were included alongside publications that made data accessible or indicated that data were available (Table 3, Categories 1 and 2), over time (Figure 6). Our findings indicate that the types of documentation required to understand and reuse research data are seldom provided in CIHR-funded publications that share data (13%, n=554). Study protocols were the most frequently included at 13.9% and data management plans were least frequent at 0.1%. Although documentation supporting reuse was scarce, the practice of including data-related documentation has grown in the past three years, with our results showing the increasing availability of data analysis plans, software code, and data collection instruments.

The CIHR-funded data sharing landscape

Institutions associated with the CIHR-funded publications included in our sample are shown in Figure 7A and 7B. The University of Toronto had the greatest proportion of CIHR-funded publications (22.18%, n=919). Among institutions with more than 10 publications, those with the greatest proportion of publications where data were accessible (Table 3, Category 1) or available (Table 3, Category 2) were the Structural Genomics Consortium (95.2%, n=20) and the University of Waterloo (48%, n=12), respectively.

Of the 4,144 publications, the journals used most frequently were PLOS One (17.76%, n=736) followed by the Journal of Biological Chemistry (5.02%, n=208). Among journals with more than 25 CIHR-funded published publications, the top three journals that included examples of accessible data (Table 3, Category 1) were the Journal of Molecular Biology, (96.6%), Molecular Cell (95.7%), and Proceedings of the National Academy of Sciences, U.S.A. (94.8%). Where data were available (Table 3, Category 2), the International Journal of Behavioral Nutrition and Physical Activity (88.46%) and BMC Medical Research Methodology (73.9%) were the most prominent (Figure 8A and 8B).

Interpretation

This study highlights significant room for growth in improving the discoverability, accessibility, and usability of CIHR-funded research data. While, encouragingly, repositories were the most common venues authors chose to share their data (37%), the remainder of shared data was made available within the publication or its supplementary files, by request or application, via a website, or by a long tail of other means (see analysis report on OSF) (26). Sharing data in a repository is recognized as best practice because it provides infrastructure for discovery, structured metadata, and long-term, reliable access. The other sharing methods we identified present by comparison significant barriers to discovery, interpretability, and access, in that they make data difficult to find, do not incorporate metadata, and do not facilitate data access consistently. These characteristics conflict with expectations outlined in Canada's Tri-Agency data management (22) and international FAIR guiding principles (15).

Metadata is an essential component of data sharing that provides valuable context about the nature of data, how they were collected, and how they can be reused (31). Metadata also improve discoverability by applying structured descriptors to data that allow them to be searched for and retrieved. Most sharing methods we encountered during our study did not incorporate metadata, making data difficult to locate, interpret, and reuse. Inadequate metadata descriptions are a recognized problem in the data sharing landscape (32–34), and the CIHR-funded data sharing practices assessed in this study are no different. Without adequate metadata, these data will remain hidden within the publication and their utility for future research remains in question (35).

In instances where authors indicated that data was available by request or application (22%), they did not provide adequate instructions on how to formally acquire the data, leaving interested researchers with no guidance on what a successful request would look like. The absence of metadata elaborating on application requirements calls into question the true availability of these data, and impedes future research based upon them. The challenges of requesting access to data in the health sciences have been studied elsewhere in relation to the inadequate transparency and standardization of data use agreements (33,36,37). Because data available by request or application are often collected from human participants, improving the discoverability of and access to these sensitive data will help prevent unnecessary study replication, create opportunities for pooling related data, and increase research efficiency to accelerate new discoveries (33,34,36,37). Our findings indicate that current practices in CIHR-funded data sharing lack the standardization and transparency necessary to secure these outcomes.

In our examination of practices that support reusability, we found that the most frequent types of documentation shared alongside data – supplementary figures and/or tables and study protocols – do not generally support the interpretation and reuse of data. Tables and figures were often reiterations of visualizations presented within the body of the publication; study protocols and data collection instruments, while helpful for contextualizing how data was gathered and analyzed, do little to help others understand the data themselves and how to

interact with them. Descriptive documentation for data such as codebooks and data dictionaries, and actionable supporting files such as code and software, are increasingly recognized as necessary components of human- and machine-readable research data (27,30,38), and our results indicate that CIHR-funded research data sharing practices can vastly improve in this area.

As Canada implements new data sharing requirements for federally funded research, our study highlights the importance of developing policies and standards at the federal and institutional levels to ensure that all research data underlying published findings have quality metadata attached to them, and that sufficient documentation to support interpretation and reuse is provided. Future directions of study should focus on the development of metadata standards for sensitive data to facilitate reuse and support transparent data request processes. We also recommend that Canadian data repositories explore how to better accommodate sensitive data so that they can be made discoverable while honouring access restrictions and privacy requirements.

Limitations of the study

This study used a sample of CIHR-funded publications pulled exclusively from PubMed and PMC. While these are the most comprehensive biomedical databases available, there are likely other databases where CIHR-funded publications exist. To manage study feasibility, we limited our review of documentation to that which was shared or stated within the publication and did not extend this analysis to repositories or websites where some research data was shared.

Conclusion

This study surveys the complex landscape of CIHR-funded data sharing practices, revealing a diverse range of data sharing methods and, in 38% of cases, an absence of data sharing altogether. It is remarkable that over 70% of publications that shared data did not incorporate sufficient metadata or documentation to facilitate discovery, access, and reuse. Without policies and standards in place that anticipate the upcoming Tri-Agency data management policy, and enhanced support for researchers seeking to implement best practices in data management and sharing, the majority of publicly funded research data will remain hidden, inaccessible, and unusable. For CIHR-funded data in particular, transparent metadata and reporting guidelines for sensitive data will be essential for improving data discoverability and accessibility across the health sciences.

Data Availability Statement

All raw, processed, and analyzed data, as well as accompanying documentation, reports and scripts are available on the Open Science Framework at https://osf.io/n9jv5.

Acknowledgements

We would like to thank the Canadian Hub for Applied and Social Research (CHASR) for their support in procuring and analyzing the research data in this study. This project was supported in part by the University of Saskatchewan Faculty Recruitment and Retention Program.

References

1. Collins, Francis. Statement on Final NIH Policy for Data Management and Sharing. National Institutes of Health. 2020.

2. FS Collins, Tabak L. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612–3.

3. Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, et al. Identifying ELIXIR Core Data Resources. F1000Research. 2016 Sep 30;5:ELIXIR-2422.

4. Artini M, Atzori C, Bardi A, Bruzzo SL, Manghi P, Mannocci A. The OpenAIRE Literature Broker Service for Institutional Repositories. -Lib Mag. 2015 Dec 11;21(11/12):95–104.

5. OpenAIRE. OpenAIRE joins forces with Canada's federal granting agencies and CARL [Internet]. Vol. 2019, OpenAIRE. OpenAIRE; 2019. Available from: https://www.openaire.eu/openaire-joins-forces-with-canada-s-federal-granting-agencies-and-carl

6. Humphreys GS, Tinto H, Barnes KI. Strength in Numbers: The WWARN Case Study of Purpose-Driven Data Sharing. Am J Trop Med Hyg. 2019;100(1):13–5.

7. McAlister VC, Harvey EJ. The benefits and risks of requiring researchers to share data. Can J Surg. 2016;59(6):364–5.

8. Owens B. Data sharing: Access all areas. Nature. 2016;533(7602):S71-2.

9. Mendelson DS, Bak PRG, Menschik E, Siegel E. Informatics in radiology: image exchange: IHE and the evolution of image sharing. Radiogr Rev Publ Radiol Soc N Am Inc. 2008 Dec;28(7):1817–33.

10. Steele Gray C, Barnsley J, Gagnon D, Belzile L, Kenealy T, Shaw J, et al. Using information communication technology in models of integrated community-based primary health care: learning from the iCOACH case studies. Implement Sci IS. 2018 26;13(1):87.

11. Vestrup JA, Phang PT, Vertesi L, Wing PC, Hamilton NE. The utility of a multicenter regional trauma registry. J Trauma. 1994 Sep;37(3):375–8.

12. Dumontier M, Wesley K. Advancing Discovery Science with FAIR Data Stewardship: Findable, Accessible, Interoperable, Reusable. Ser Libr. 2018;74(1–4):39–48.

13. Reiser L, Harper L, Freeling M, Han B, Luan S. FAIR: a call to make published data more findable, accessible, interoperable, and reusable. Mol Plant. 2018;11(9):1105–8.

14. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251).

15. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3.

16. Persaud N. A national electronic health record for primary care. CMAJ Can Med Assoc J J Assoc Medicale Can. 2019 Jan 14;191(2):E28–9.

17. Rathi VK, Strait KM, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. Trials. 2014 Oct 2;15:384.

18. Government of Canada. Draft Tri-Agency Research Data Management Policy for Consultation [Internet]. Government of Canada Policies and Guidelines. 2018. Available from: http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html

19. Kelsall D. New CMAJ policy on sharing study data. CMAJ Can Med Assoc J. 2017;189(34):E1082–E1082.

20. Wilson L. Exploring the Canadian Federated Research Data Repository Service. Biodivers Inf Sci Stand. 2017;

21. Government of Canada. Digital Research Infrastructure [Internet]. Government of Canada Innovation, Science, and Economic Development Canada. 2019. Available from: http://www.ic.gc.ca/eic/site/136.nsf/eng/home

22. Government of Canada. Tri-Agency Statement of Principles on Digital Data Management [Internet]. Government of Canada Policies and Guidelines. 2016. Available from: http://science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html

23. National Library of Medicine. Data Filters in PMC and PubMed [Internet]. NLM Technical Bulletin. 2018. Available from: https://www.nlm.nih.gov/pubs/techbull/ma18/brief/ma18_pmc_data_filters.html

24. CASRAI. Research data definition [Internet]. CASRAI Glossary. 2019. Available from: https://casrai.org/term/research-data/

25. National Library of Medicine. Open Access Subset [Internet]. PMC Tools. 2019. Available from: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

26. Read K, Ganshorn H, Rutley S, Scott D. Surveying the landscape of CIHR-funded research data sharing practices: An analysis of the published literature [Internet]. Open Science Framework; 2020. Available from: https://osf.io/n9jv5

27. Bakken S. The journey to transparency, reproducibility, and replicability. J Am Med Inform Assoc. 2019;26(3):185–7.

28. Holub P, Kohlmayer F, Prasser F, Mayrhofer MT, Schlünder I, Martin GM, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. Biopreservation Biobanking. 2018 Apr;16(2):97–105.

29. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain. 2020;13(1):24–24.

30. Walters WP. Code Sharing in the Open Science Era. J Chem Inf Model. 2020 Oct 26;60(10):4417–20.

31. Qin J, Ball A, Greenberg J. Functional and Architectural Requirements for Metadata. In: International Conference on Dublin Core and Metadata Applications. 2012. p. 10.

32. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform. 2020 Jul;107:103421.

33. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, et al. Barriers to accessing public cancer genomic data. Sci Data. 2019 Jun 20;6(1):98.

34. Vassar M, Jellison S, Wendelbo H, Wayant C. Data sharing practices in randomized trials of addiction interventions. Addict Behav. 2020 Mar;102:106193.

35. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data. 2020;7(1):5–8.

36. Miller J, Ross JS, Wilenzick M, Mello MM. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. BMJ. 2019 Jul 10;366:l4217.

37. Shabani M, Obasa M. Transparency and objectivity in governance of clinical trials data sharing: Current practices and approaches. Clin Trials Lond Engl. 2019 Oct;16(5):547–51.

38. Davenport JH, Grant J, Jones CM. Data Without Software Are Just Numbers. Data Sci J. 2020;19(1).

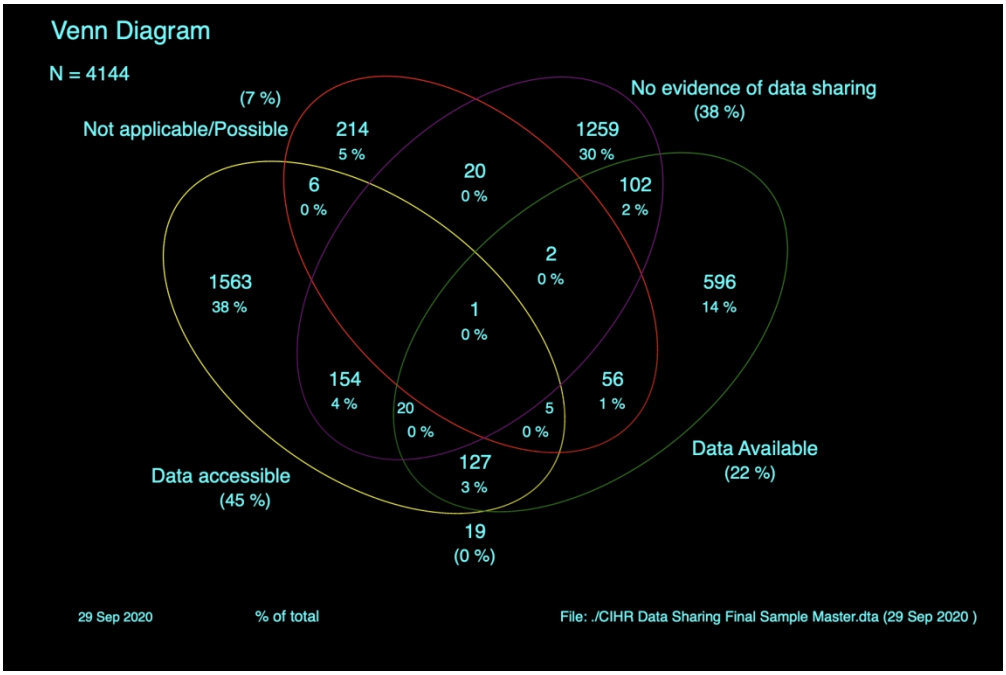


Figure 1. Frequency of publications categorized by data sharing status (n=4,144)

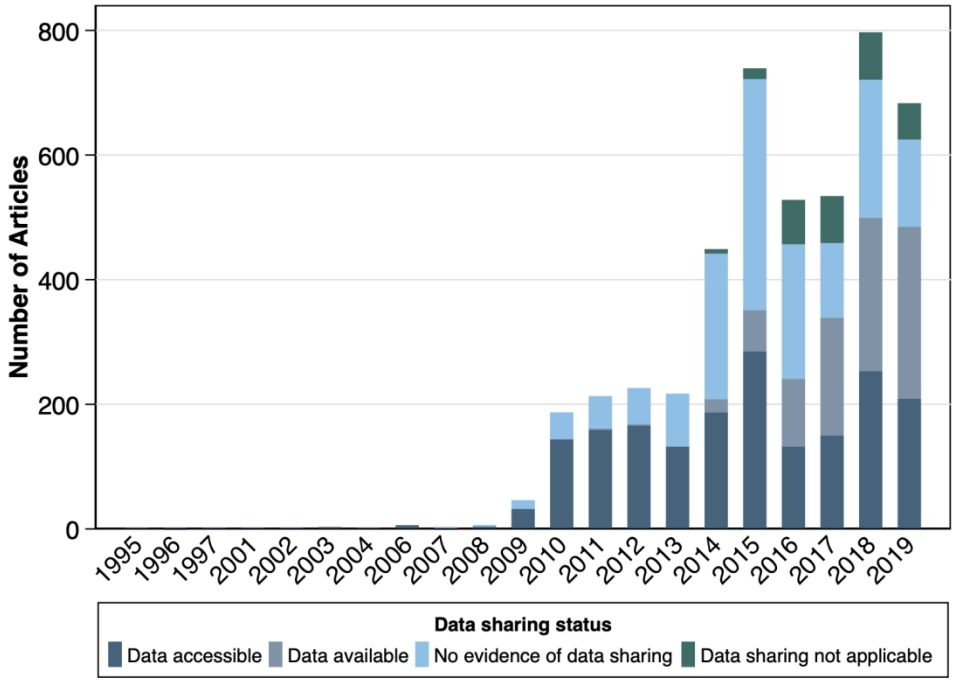


Figure 2. Frequency of data sharing status category over time -- not mutually exclusive to a single publication (n=4,144)

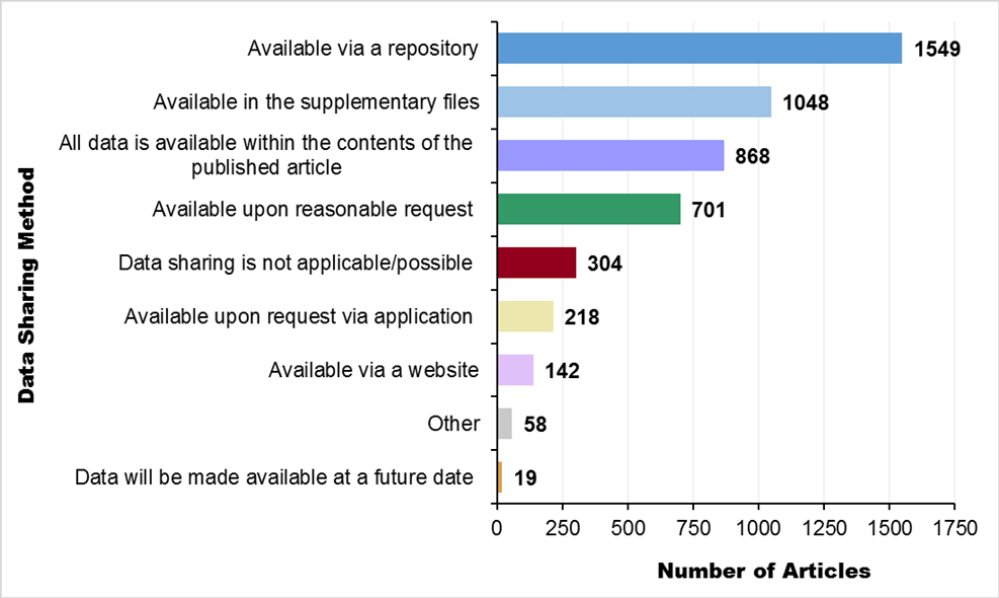


Figure 3. Frequency of data sharing method -- not mutually exclusive to a single publication (n=4,144)

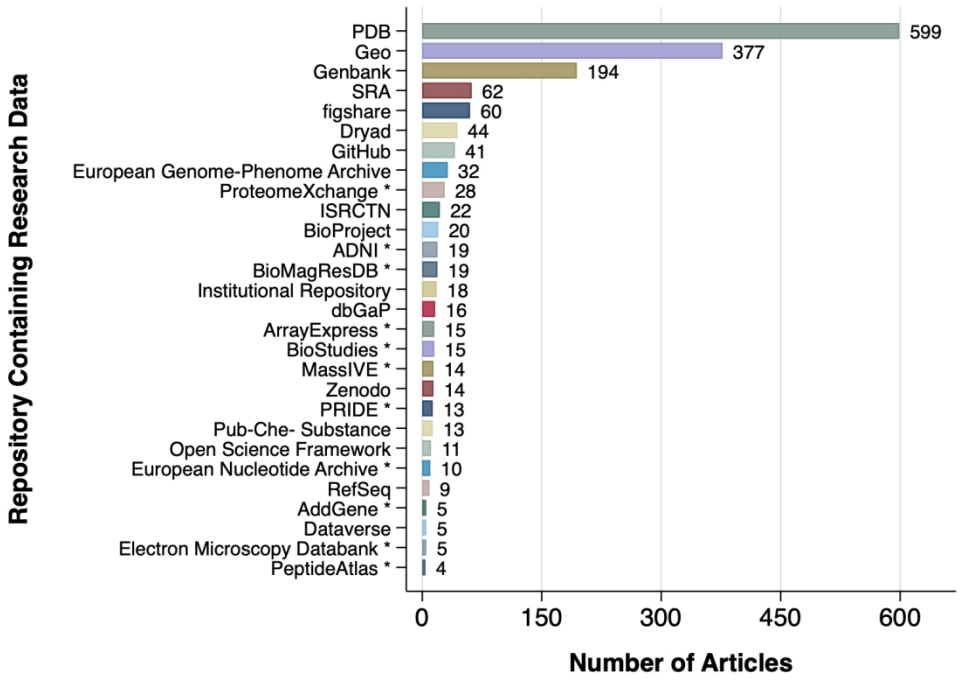


Figure 4. Frequency of data repositories used to store CIHR-funded research data (n=1,544). Asterisks denote repositories that were not anticipated within our instrument, and were added from a free-text "Other" category during data collection.

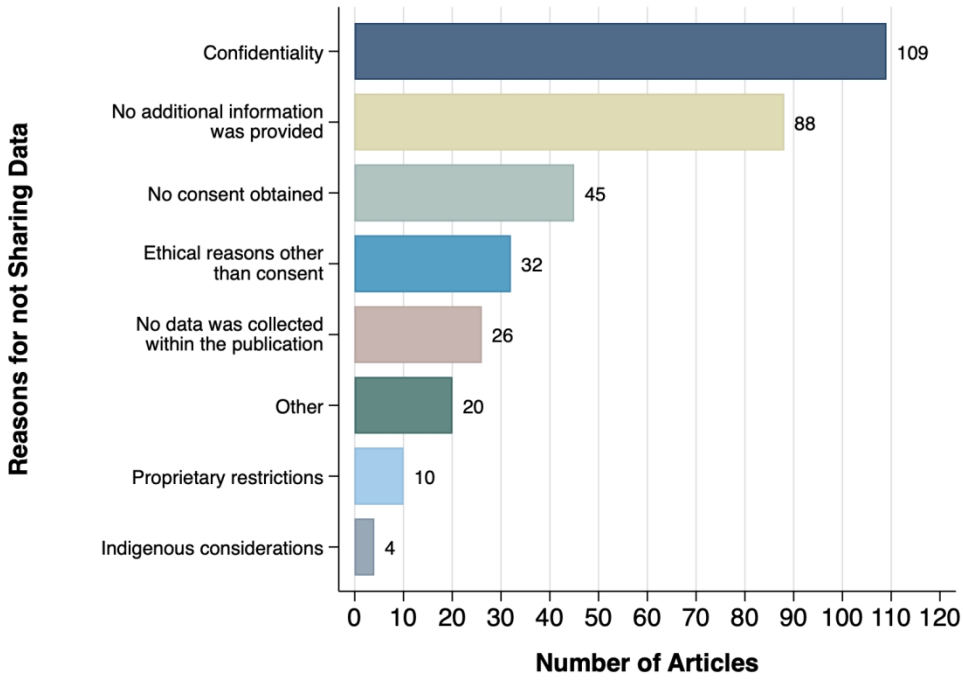


Figure 5. Frequency of reasons for not sharing data (n=300)

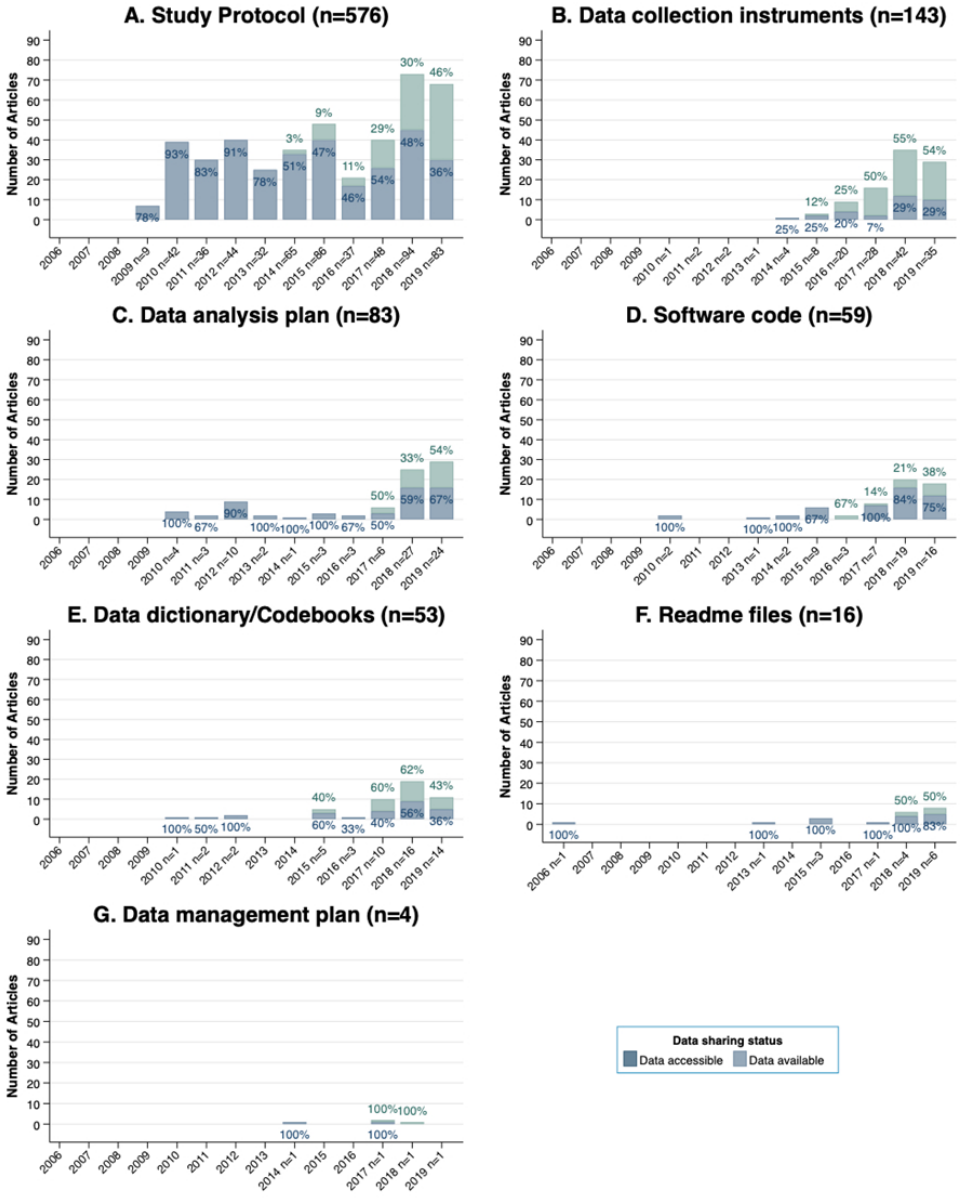


Figure 6. Frequency of data management-specific documentation grouped by data sharing status over time (n=4,144)

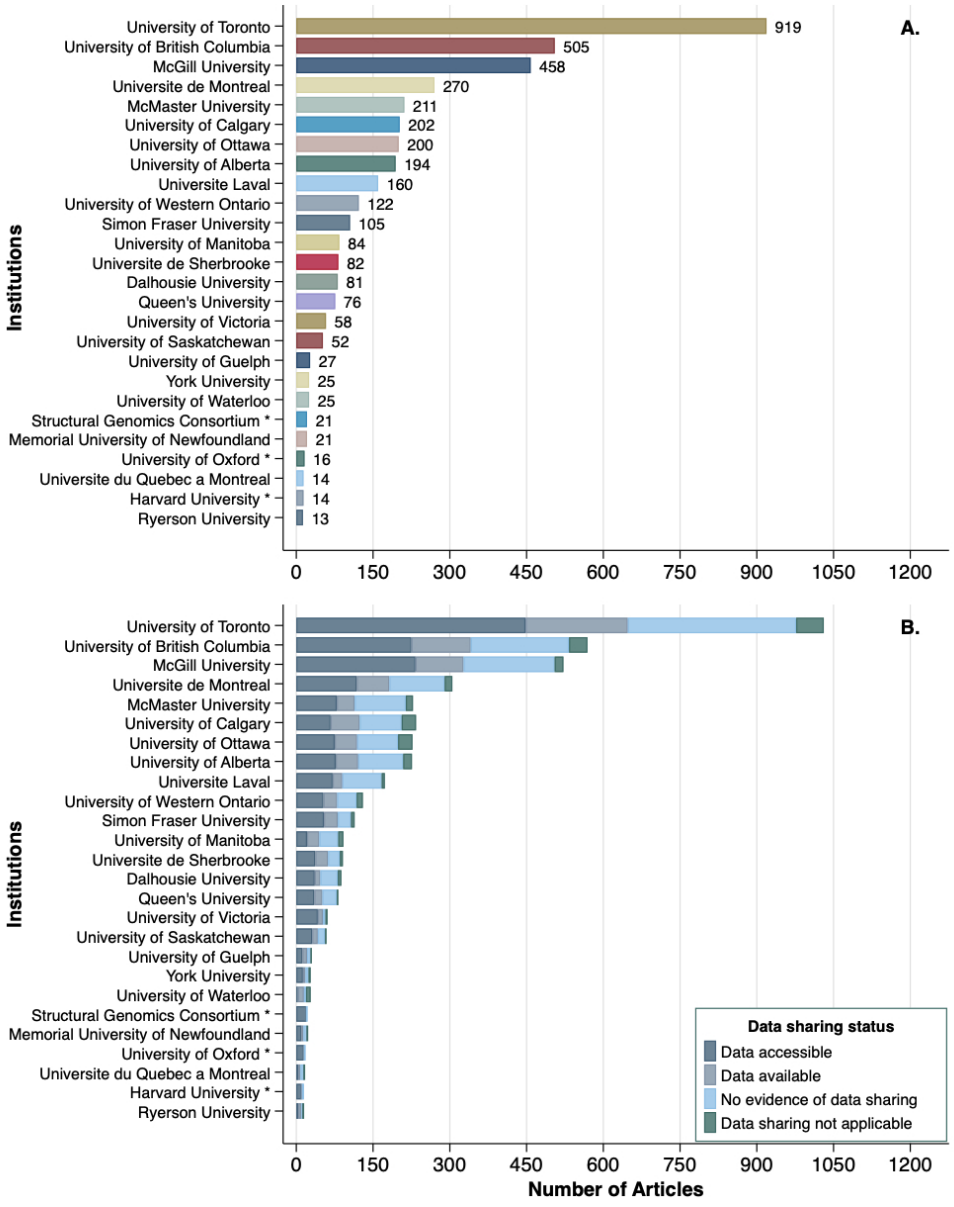


Figure 7A. Frequency of publications identified by institution (n=4,144)
Figure 7B. Frequency of publications identified by institution grouped by data sharing status (n=4,144)

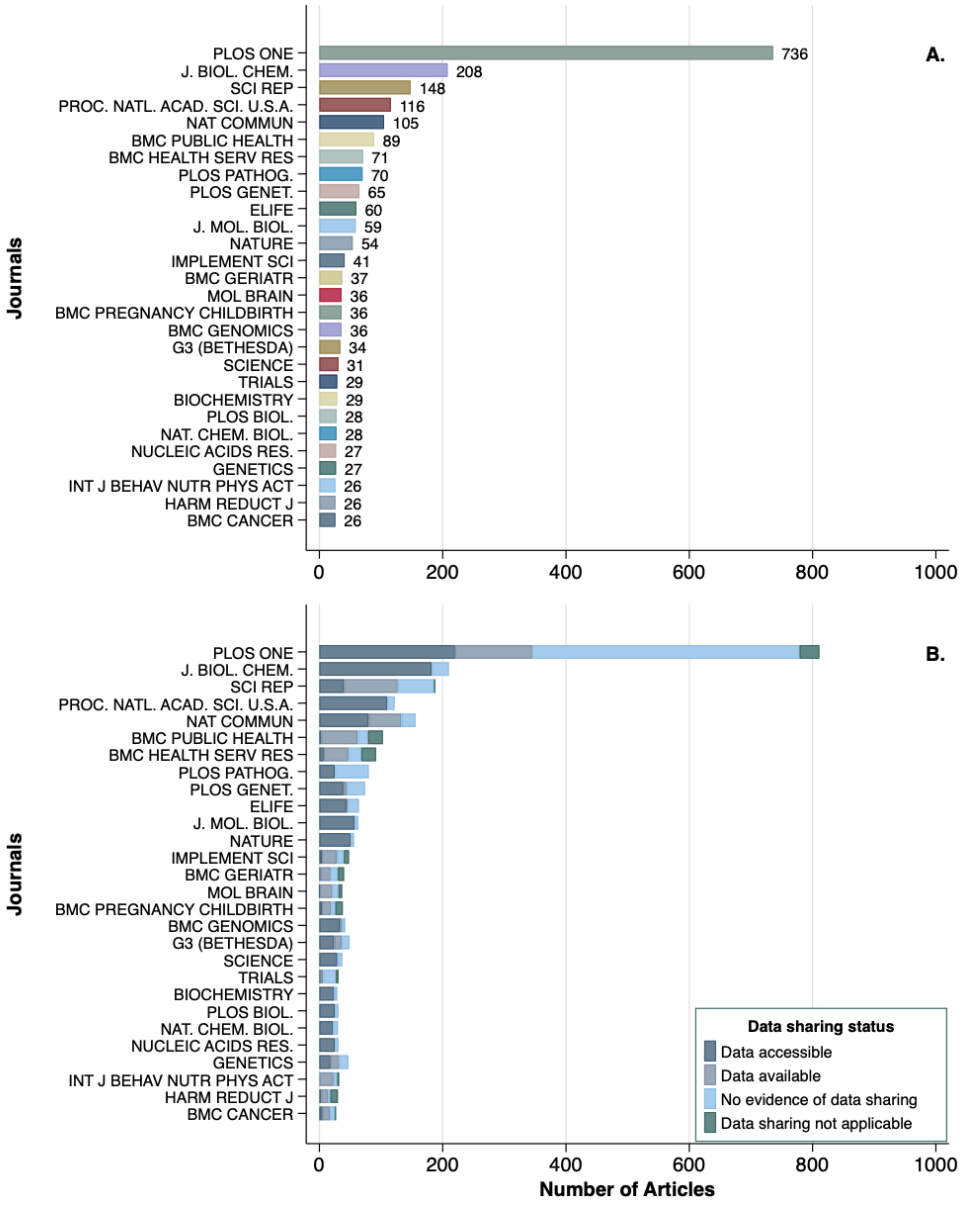


Figure 8A. Frequency of publications identified by journal (n=4,144)
Figure 8B. Frequency of publications identified categorized by data sharing status (n=4,144)

Table 1. Search strategy per database

Database	Search Filter	CIHR-strategy	Results
PubMed Central	“has associated data”[filter] OR “has data citations”[filter]	("canadian institutes of health research"[Grant Number] OR cihr[grant number] OR IRSC[grant number] OR “Instituts de recherche en sante du Canada”[Grant Number] OR IRSC[Grant Number])	2536
PubMed	data[filter]		2624

Table 2. Extracted metadata fields from PMC, PubMed, and the PMC Open Access Subset

Metadata Field	Description	PMC metadata	PMC Open Access Subset	MEDLINE / PubMed Dataset metadata
Author affiliation	Includes the institutional affiliation and address (including email address, when available) of the authors of the publication as it appears in the journal.	x	x	x
Publication date	The date that the publication was published.	x	x	x
Journal title	The journal title abbreviation, full journal title, or ISSN number	x	x	x
Publication type	The type of publication as categorized by MEDLINE	x	x	x
Corresponding author	The name of the corresponding author of the publication	x	x	x
Data availability statements	publications or manuscripts with data availability statements		x	
Data citations	publications or manuscripts with data citations		x	
MeSH Major Topic Headings	A MeSH term that is one of the main topics discussed in the publication.	x	x	x
publication body - Key Terms	Includes all key terms in the body of a publication except for the Abstract and References.		x	
Acknowledgements	Includes all words in the acknowledgement section of a publication (e.g., “National Institutes of Health[ack]”).	x	x	

Grant number	The grant number search field includes research grant numbers, contract numbers, or both that designate financial support by Agency of the US PHS (Public Health Service), and other national or international funding sources.	x	x	x

Table 3. Data sharing status categories and their definitions

Status Category	Definition
(1) Data accessible	Research data files could be identified, accessed, and downloaded.
(2) Data available	Authors stated either within the manuscript, the data availability statement, or the acknowledgements that research data was available upon request or via an application process.
(3) Data sharing not applicable/possible	Authors stated either within the manuscript, data availability statement, or acknowledgements that research data sharing was not possible or applicable.
(4) No evidence of data sharing	Authors made no mention of data sharing within the manuscript, data availability statement, or acknowledgements; indicated data sharing would be available at a future date; or the publication contained no research data files.

Table 4. Frequency of data sharing status -- not mutually exclusive to a single publication (n=4,144)

Data sharing method	Frequency (n)	Percent (%)
Data accessible	1,876	45.27
No evidence of data sharing	1,558	37.60
Data available	909	21.94
Data sharing not applicable or possible	304	7.34

Table 5. Documentation identified and categorized by data sharing status

Type of Information	Data sharing status (by count)			
	Data Accessible	Data Available	No evidence of data sharing	Data sharing not possible/applicable
Supplementary figures and/or tables	1,258	385	881	65
Study protocol	332	94	200	18
Data files	504	82	3	10
Transparent reporting form	138	57	107	4
Data collection instruments	31	62	52	39
Videos	56	20	41	1
Data analysis plan/documentation	58	25	12	4
Image files	23	5	43	0
Software code	46	13	8	2
Data dictionary/codebook	26	24	10	6
Preservation formats for structured data	50	4	2	0
Readme files	15	5	1	1
Data management plan	2	2	0	0
Other	29	30	28	6