MLA HYPOTHESIS
Journal of the Research Caucus

Sharing research data to comply with a journal policy: Experience of a first-time depositor

Marianne D. Burke PhD, MA, AHIP
Associate Professor Emerita
University of Vermont
Burlington, VT

ABSTRACT

Background Journals in health sciences increasingly require or recommend that authors deposit the data from their research in open repositories. The rationale for publicly available data is well understood but many researchers lack the time, knowledge, and skills to do it well, if at all. There are few descriptions of the pragmatic process a researcher author undertakes to complete the open data deposit in the literature. When the author's manuscript for a mixed methods study was accepted by a journal that required shared data as condition of publication, she proceeded to comply despite uncertainty with the process.

Purpose The purpose of this study is to describe the experience of an information science researcher and first-time data depositor to complete an open data deposit. The study illustrates the questions encountered and choices made in the process.

Methods To begin the data deposit process, the author found guidance from the accepting journal's policy and rationale for its shared data requirement. A checklist of pragmatic steps from an open repository provide a framework that the author used to outline and organize the process. Process steps included organizing data files, preparing documentation, determining rights and licensing, and determining sharing and permissions. Choices and decisions included which data versions to share, how much data to share, repository choice, and file naming. Processes and decisions varied between the quantitative and qualitative data prepared.

Results The author deposited data in two datasets and documentation for each in Figshare open repository, thus meeting the journal policy requirements to deposit sufficient data and documentation to replicate the results reported in the journal article and also meeting the publication deadline to include a Data Availability Statement with the published article.

Conclusion This experience illustrated some practical data sharing issues faced by a librarian author seeking to comply with a journal data sharing policy requirement for publication of an accepted manuscript. Both novice data depositors and data librarians

may find this individual experience useful for their own work and the advice they give to others.

BACKGROUND

Journals in health sciences increasingly require or recommend that authors deposit the data from their research in open repositories. In October 2019, the Journal of the Medical Library Association (JMLA) announced its policy to require authors to deposit deidentified research data for all original investigations and case report articles accepted for publication in October of 2019 [1]. Although many library and information science journals recommend data sharing, JMLA was the first and still possibly the only one to require it.

The rationale for publicly available data—that it fosters scientific progress and enables replication and reproducibility of research—is well described in the literature and in the JMLA editorial justifying its requirement [2]. Data sharing workflows for researchers and for repositories are described at a conceptual level by Austin et al [3] with little process detail. Educational methods for teaching research data management methods to students and researchers are described in a review by Corti and Van den Eyden. [4]. They recommend skills training and active learning methods but do not cover specific content on how to share data and what data to share.

While most researchers support data sharing and agree with its value, studies show that many researchers have uncertainty and concerns with the pragmatic aspects of data sharing [5-7]. A 2018 survey by Stuart et al of 7,000 researchers from various disciplines and experience levels found that 76% rated the importance of discoverable data highly, while 46% reported organizing data and presenting data as problematic. Lack of time was a problem for 35%, and repository choice a problem for 33% [8]. A survey of clinical trial investigators, by Tannenbaum et al, found that they spent a median of 18 hours (IQR 8 – 40) per data set shared [9]. A meta-synthesis of qualitative studies of researcher experience with data sharing found that “researchers lack time, resources and skills to effectively share their data in public repositories”[10].

I submitted a manuscript to *JMLA* after its data sharing policy was announced, but prior to when it went into effect. Technically I was not required to share the data for manuscript. But when my manuscript was accepted for the July 2020 issue, I decided to share the data from my research motivated by belief in the value of data sharing and by the journal requirement policy. Like other researchers, I was uncertain about how to organize and prepare the data, and was apprehensive about the time it would take, but I decided to do it.

The accepted manuscript (now a *JMLA* article) is a mixed methods assessment of a clinical evidence technology based on a survey of 32 primary care providers (PCPs) who participated in an earlier randomized trial concerning their use (or non-use)

of the technology during that trial, and interviews with 11 PCPs in the intervention arm [11].

While advantages and obstacles to researcher data sharing are well described, few studies describe researchers' data sharing experiences in detail. The purpose of this case study is to describe my experience as an information science researcher and first-time data-depositor, and illustrate the practical questions and issues that I encountered in the process.

METHODS

To begin the open repository deposit process, I reviewed the journal's requirements. The *JMLA* data sharing policy states "The *JMLA* requires authors of Original Investigation, Case Report, and Special Paper articles to (1) place the de-identified data associated with the manuscript in a repository and (2) include a Data Availability Statement in the manuscript describing where and how the data can be accessed" [1]. The *JMLA* editorial announcing the policy by Akers et al elaborated on the requirements stating that "at least minimal data needed to support or replicate results", and "documentation describing the contents of the data files" must be deposited [2]. A data availability statement, including a URL or DOI for the data, must be provided by the author and included with the published article.

The editorial offered guidance for authors depositing data for the first time and recommended a selection of resources and references that could be consulted for help, including Library-based web-sites and LibGuides, and webpages of open repositories and noted that help is available from health science libraries and librarians [2]. I sought practical guidance that provided specific data sharing information and advice, and not data sharing background or rationale from among the editorial references. Three references that provided pragmatic methods for organization and presentation of data were Washington University Libraries' webpage "Preparing data for deposit" [12], the 2019 Journal of eScience Librarianship Presentations article entitled "Best Practices for Data Sharing and Deposit for Librarian Authors", by Regina Raboin et al [13] , and the Inter-University Consortium for Political and Social Research (ICPSR) Guide which includes a section entitled "Preparing data for sharing" [14].

I found clear steps to follow in the Digital Research Materials Repository (DRMR) Deposit website created by the Washington University St. Louis Libraries (available at https://libguides.wustl.edu/drmr/dataprep) [12]. There, a checklist and a template README file form are available as downloadable and printable tools to address the organization, presentation, documentation and other tasks. The Checklist outlines four major steps and multiple itemized sub-steps.

The WUSL Data Research Materials Repository (DRMR) checklist tasks are to organize your deposit, prepare documentation, determine deposit rights & licensing, and determine sharing and permissions. I used the DRMR checklist as a process framework

to move forward with the data deposit but revised it adding “review journal policy and requirements” as first step described above and a fifth task, “choose your data repository”. The process was recursive in that I worked on several checklist steps at once or went back and forth between them rather than proceeding in a consecutive workflow. [See Figure 1]

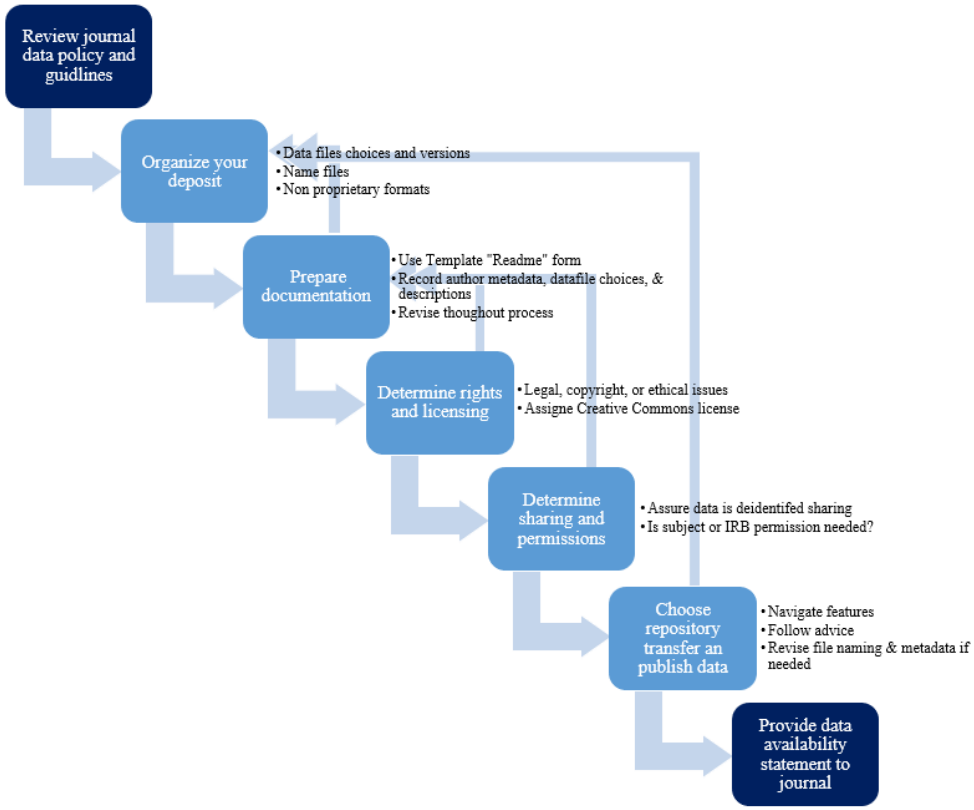

Figure 1: Author's process to deposit research data in an open repository to comply with journal policy.

Step 1: Organize Your Deposit

After reviewing the journal’s policy and guidelines, the first step, to “Organize your deposit” seemed simple, but it was on this step that I encountered the most decision points and spent the most time. Sub-items in this category included gathering and choosing the dataset files to share, naming files and labelling variables consistently, and transforming datasets and files to -non-proprietary file formats.

The *JMLA* policy names appropriate data types as “including but not limited to spreadsheets, text files, interview recordings or transcripts, images, videos, output from statistical software, and computer code or scripts”[2]. I had several of these data types and felt uncertainty concerning which versions of similar data files and how much of the total available data to prepare.

Step 1a: Data version

Owing to the mixed methods design of the research, there were both quantitative and qualitative data files, and version was an issue in each type. For the quantitative data, potential data files included raw deidentified survey response data output from REDCap [15], and the same survey data cleaned of extraneous variables and processed for analysis from the Stata statistical software [16]. Raboin et al noted that either raw or processed data may be used [13] but recommend REDCap export data because the system can automatically remove personal identifying information and provides a codebook or data dictionary that describes all variables. On the other hand, processed data, i.e. data cleaned of extraneous or administrative variables and possibly containing new variables generated and used in analysis, may more completely represent the reported results. *JMLA* policy did not address this difference, but PLoS guidelines state, "Authors do not need to submit the raw data collected if the standard in the field is to share data that have been processed."[17]. I did not know of a library science standard but, in Clinical and Translational Science, where I was a graduate student, the usual practice was to import deidentified raw data into the statistical package (Stata) and clean it and sometimes generate new variables. For this reason, I chose the processed Stata dataset to deposit. In addition, I edited the Data Dictionary exported from REDCap and included it as a deposited file.

For the qualitative data, there were two versions of the interviews: digital recordings and full transcripts. The digital recordings were difficult to edit and de-identify given my lack of technical skill and the possibility of voice recognition of the participants. I decided to deposit the deidentified transcripts, with the caveat they would need further review for deidentification and anonymization (see checklist step 4). I also considered preparation of the fully coded interview transcripts as analyzed in NVivo [18] for deposit. I found NVivo exported tables in a word processing format (.docx) difficult and time-consuming to edit and convert into a text (.txt) document. In addition, this information would be redundant since many exemplar quotations and associated codes were included as supplementary material in the accepted manuscript. Therefore, I decided not to deposit the complete coded interviews.

Step 1b: Entire data or portion

Whether to share all existing data results or just those reported in the accepted manuscript was a decision point. According to the *JMLA* policy "authors are expected to deposit at least the minimum amount of data needed to reproduce the results described in the manuscript" [1]. A similar policy by PLoS states, "Authors do not need to submit their entire data set if only a portion of the data were used in the reported study" [17]. I considered depositing data from a baseline survey analyzed and reported in an earlier related journal article, but not yet shared [19]. The baseline data were not used or analyzed in the results in the current accepted manuscript. Depositing the complete

data might be considered a better practice because the data are available sooner for review and secondary analyses. However, I decided to deposit only the data needed to reproduce the results of the manuscript to reduce the organization and documentation effort and save time. I plan to share all data from the earlier study in the near future.

Another task at the "Organize your deposit" step was to convert all data files to non-proprietary formats. I did this for the chosen files without much difficulty. Tabular survey data was exported from Stata as comma separated values (.csv) files, and the interview transcripts were converted from word processing documents (.docx) to text (.txt) files for the deposit.

The last task at this step was to create consistent labels for data headers, files, and variables. For me, dataset file naming was an iterative process and was only concluded with the final deposit. Variable names and data column headers in the quantitative data were consistent as exported from Stata and REDCap as tabular data.

Step 2: Prepare Data Documentation

This step is the preparation of the metadata, the data descriptions of each data file, the names of files deposited and so on. The downloadable WUSL README (https://libguides.wustl.edu/ld.php?content_id=29108244) template form described earlier presented a complete list of information elements needed to be provided with the data. The form kept track of progress on decisions at each of the outlined steps creating an inventory of file choices, naming and documentation. Information recorded on this form was drawn from every other step in the process and served as a complete source of documentation for the deposit.

Step 3: Determine Rights and Licensing

This step concerned the rights of others to reuse of the data. Since this would be an open data deposit, I knew that a Creative Commons (CC) license was appropriate, but there are several types of CC licenses so this needed to be checked. License types are defined at https://creativecommons.org/licenses/. I chose the CC BY license which lets others distribute remix and adapt your data as long as they credit the author for the original creation. I did not seek any copyright permissions for the deposit because my co-authors and I were the creators of the data.

Step 4: Determine Sharing and Permissions

This step involved assuring that authorization and consent to share data collected from participants in a study was obtained. For the study reported in the accepted manuscript, participants explicitly consented to their data being deidentified, aggregated, analyzed, and reported in a research journal in an IRB approved consent form. Deidentification of participants in the quantitative data was easily settled with the

automated removal of personal information step that an investigator utilizes when exporting data from REDCap.

Deidentification in qualitative data is more problematic [14]. Studies have shown that even when data is deidentified by the removal of personal information, deductive disclosure i.e. logical supposition of who the source is by members of the same community is possible, so additional anonymization is needed [20]. I had not explicitly stated the likelihood of deposit of full transcripts of interviews in the consent information for provider participants, only that their answers to questions would be aggregated with those of others. I sent a sample transcript to the University of Vermont Human Protections Office for guidance. After review, the IRB administrator ruled that transcripts deposit would be allowed because there were no direct identifiers and the "information would not be considered sensitive in nature." i.e. they posed little risk of harm to the primary care provider participants [Email communication to Author from L. Crain "Re: Research Data Sharing in Repositories", 31 Jan 2020]. I chose to deposit the transcripts after removing additional possibly identifying clues such as clinic site and idiosyncratic phraseology.

Step 5: Choose Repository

For repository choice, I found criteria to consider in the Raboin presentation (slide 26) [13], Journal recommendation, visibility through a DOI, usability, and features such as embargoes, creative commons licensing, and formats accepted were meaningful to me. Several general non-discipline specific open repositories were suggested by *JMLA*. I reviewed a comparison grid of general data repositories at a Harvard Medical School website that included four of the journal-recommended repositories: Dataverse, Dryad, Figshare and Zenodo [21]. A librarian colleague who had deposited data, advised that Figshare [22] was easy to use and provided prompts and mini-tutorials. Without having a strong preference, I chose Figshare. When exploring the Figshare website, I found simple instructions and few restrictions. I used trial and error methods to learn the repository's features and began uploading the data files and documentation. For each dataset entered in Figshare, there are prompts for metadata including title, authors categories, links to related keywords, data description, any restrictions on use of data, and sharing level. My prepared README form contained most of the required information so I cut and paste the authors, titles, descriptions, file names, license and permissions into the Figshare metadata boxes. Figshare offered additional advice on naming and other metadata conventions to increase findability and I did rename files and datasets to increase findability. I entered separate documentation metadata for the qualitative and quantitative datasets based on my README form. I named dataset authors based on who had participated in design and review of each set, and each dataset was described differently [23, 24]. I did

change some metadata and file names after publishing. This resulted in the version numbers being added to the DOIs.

RESULTS

I deposited data in two separate datasets in Figshare, one for the quantitative survey dataset, https://doi.org/10.6084/m9.figshare.11893875.v3 and one for the qualitative https://doi.org/10.6084/m9.figshare.11893956.v1. The survey dataset contains metadata description, tabular processed data with 39 variables and 22 observations, and a data dictionary file that includes the variable names, corresponding survey questions, and response type and possible responses in a comma separated file. The qualitative dataset contains metadata description and the interview transcripts as a single text (.txt) data file.

I spent at least 30 hours over three weeks reading polices, referring to links in library LibGuides, conferring with my institution's data librarian and researcher colleagues, and performing the tasks in the checklist outline, making decisions, and navigating the Figshare repository. I met the journal policy requirements to deposit sufficient data and documentation to reproduce the findings of the manuscript, and met the publication deadline for including the Data Availability Statement with the published article.

DISCUSSION

My experience supports the finding of Kim and Stanton that authors' belief in the value and benefits of data sharing combined with the open data sharing requirement of journals may motivate authors to share data, but the time and effort needed to complete the process may negatively impact the researchers' actual participation in data sharing [5].

As anticipated, my data deposit experience was time-consuming and presented myriad decision-points without exact answers even with consulted sources. However, this was a first-time experience; I suspect that there will be fewer obstacles next time, although different data types and file sizes will certainly present different hurdles. I will utilize checklist and pragmatic tools again, and will be less hesitant to make decisions.

My advice to researcher/authors who are novice data sharers is to use the most specific checklists and tools you can find or make your own. Pragmatic advice and tools found on the websites supporting open data repositories such as the WSU, ICPSR and Figshare repositories may be the most useful. I believe it is better to publish the data even if some uncertainties linger. Don't let hesitation or imperfection prevent completion of the deposit. It is possible to revise metadata and upload and deposit revised data versions in repositories if that is needed.

Implications for health science librarians

Support for the data sharing activities needs of researchers may not be fully developed in health sciences libraries. While many academic health science libraries provide some level of data services, according to a survey by Ragon of academic health science library directors, 25% of those surveyed did not provide data services in their libraries and did not plan to do so in the future [25]. A survey of data librarians by Federer revealed that most of them spend less than half of their work time on data services [26]. A taxonomy of skills needed by data librarians proposed in the same article did not include skills or experience in data sharing itself. I appreciated the advice and recommendations from the library website and the data librarian at my institution, but they answered few of my specific data sharing questions. Library data services could emphasize pragmatic data sharing tools and tips at least as much as background and advocacy information on their research data management web-sites. Library-sponsored teaching sessions that utilize group workshop formats and in-class case presentations of data sharing by researchers could provide a more practical learning experience for researcher authors. Data librarians with experience sharing their own data might be better equipped to assist novice depositors (including library and information science researchers) with the problems and decision-points that arise in the process of data sharing.

Limitations

This individual case study reflects only one data sharing experience. Other researcher authors will find different issues, choices, and take-aways in their experience particularly if they are sharing data from randomized trials, other empirical designs, large data sets, or complex statistical analyses. Nevertheless, this description of my experience may give some insight to the process overall.

CONCLUSIONS

The burden placed on researcher authors when depositing data is significant. Having experienced the process once, I now see where pitfalls and problems make open data sharing problematic for researcher authors. As more health science librarians gain experience depositing their own research data, they may provide more concrete and practical answers when assisting researcher/authors and thereby reduce author uncertainty and time spent in the data sharing process.

This experience illustrated some practical data sharing issues faced by a librarian author seeking to comply with a journal data sharing policy requirement for publication of an accepted manuscript. Both novice data depositors and data librarians may find this individual experience useful for their own work and the advice they give to others.

REFERENCES

1. Journal of the Medical Library Association Data Sharing Policy: Journal of the Medical Library Association; 2019 [May 20 2020]. Available from: http://jmla.mlanet.org/ojs/jmla/about/editorialPolicies#custom-0.
2. Akers KG, Read KB, Amos L, Federer LM, Logan A, Plutchak TS. Announcing the Journal of the Medical Library Association's data sharing policy. J Med Libr Assoc. 2019 Oct;107(4):468-71. DOI: http://dx.doi.org/ 10.5195/jmla.2019.801
3. Austin CC, Bloom T, Dallmeier-Tiessen S, Khodiyar VK, Murphy F, Nurnberger A, et al. Key components of data publishing: using current best practices to develop a reference model for data publishing. International Journal on Digital Libraries. 2017 2017/06/01;18(2):77-92. DOI: http://dx.doi.org/ 10.1007/s00799-016-0178-2
4. Corti L, Van den Eynden V. Learning to manage and share data: jump-starting the research methods curriculum. International Journal of Social Research Methodology. 2015 2015/09/03;18(5):545-59. DOI: http://dx.doi.org/ 10.1080/13645579.2015.1062627
5. Kim Y, Stanton JM. Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. Journal of the Association for Information Science and Technology. 2016; 67:776–99. DOI: http://dx.doi.org/ http://dx.doi.org/10.1002/asi.23424
6. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS One. 2020;15(3):e0229003. DOI: http://dx.doi.org/ 10.1371/journal.pone.0229003
7. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. PLoS One. 2015;10(6):e0129506. DOI: http://dx.doi.org/ 10.1371/journal.pone.0129506
8. Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft M, et al. Practical challenges for researchers in data sharing. Springer Nat DOI: https://doi org/106084/m9 figsh are. 2018;59750(11):v1
9. Tannenbaum S, Ross JS, Krumholz HM, Desai NR, Ritchie JD, Lehman R, et al. Early experiences with journal data sharing policies: a survey of published clinical trial investigators. Ann Intern Med. 2018 Oct 16;169(8):586-8. DOI: http://dx.doi.org/ 10.7326/m18-0723
10. Perrier L, Blondal E, MacDonald H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. PLoS One. 2020;15(2):e0229182. DOI: http://dx.doi.org/ 10.1371/journal.pone.0229182
11. Burke MD, Savard LB, Rubin AS, Littenberg B. Barriers and facilitators to use of a clinical evidence technology in the management of skin problems in primary care: insights from mixed methods. J Med Libr Assoc. 2020 Jul 1;108(3):428-39. DOI: http://dx.doi.org/ 10.5195/jmla.2020.787

12. Digital research materials repository: preparing data for deposit [Internet].Washington University St. Louis. The Libraries. [cited 20 Aug 2020]. Available from: https://libguides.wustl.edu/drmr/dataprep.
13. Raboin R, Plutchak T, Palmer L, Goldman J. Best practices for data sharing and deposit for librarians. Journal of eScience Librarianship Presentations [Internet]. 2019; 2020 (May 2020). Available from: https://escholarship.umassmed.edu/jeslib_presentations/1
14. Guide to Social Science Data Preparation and Archiving. Phase 5 preparing data for sharing [Internet]. ICPSR: Inter-university Consortium for Political and Social Research [cited 2020_8_20]. Available from: https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.
15. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81. DOI: http://dx.doi.org/10.1016/j.jbi.2008.08.010
16. StataCorp: Stata Statistical Software: Release 14 College Station, TX: StataCorp LP; 2015.
17. Data Availability [Policy]: PLoS Journals; 2019 [updated December 5, 2019; cited 2020 2020_3_27]. Available from: https://journals.plos.org/plosone/s/data-availability.
18. NVivo Qualitative Data Analysis Software Version 12 ed: QSR International Pty Ltd.; 2018.
19. Burke M, Littenberg B. Effect of a clinical evidence technology on patient skin disease outcomes in primary care: a cluster-randomized controlled trial. J Med Libr Assoc. 2019 Apr;107(2):151-62. DOI: http://dx.doi.org/ 10.5195/jmla.2019.581
20. Tsai AC, Kohrt BA, Matthews LT, Betancourt TS, Lee JK, Papachristos AV, et al. Promises and pitfalls of data sharing in qualitative research. Soc Sci Med. 2016 Nov;169:191-8. DOI: http://dx.doi.org/ 10.1016/j.socscimed.2016.08.004
21. Harvard Biomedical Data Management: Best practices & support services for research data lifecycles: Harvard University; [updated 2018_07_062020__9_9]. Available from: https://datamanagement.hms.harvard.edu/repositories.
22. Figshare [Internet]. London: Figshare; 2013 [cited September 15, 2020]. Available from: https://figshare.com/
23. Burke M, Littenberg B. Use of a clinical evidence technology for skin disease in primary care: clinician survey data [Dataset]. figshare 2020.
24. Burke M, Savard L, Rubin A, Littenberg B. Use of a clinical evidence technology for skin disease in primary care: clinician interviews [Dataset]. figshare; 2020.
25. Ragon B. Alignment of library services with the research lifecycle. J Med Libr Assoc. 2019 Jul;107(3):384-93. DOI: http://dx.doi.org/ 10.5195/jmla.2019.595
26. Federer L. Defining data librarianship: a survey of competencies, skills, and training. 2018. 2018 2018-07-02;106(3):10. DOI: http://dx.doi.org/ 10.5195/jmla.2018.306

MLA HYPOTHESIS
Journal of the Research Caucus