Data Curation Final Report

Anne Strawbridge

University of Denver – MILS 4220

I. Data and Metadata Profile

Link to Dataset: https://datadryad.org/stash/dataset/doi:10.7272/Q6ZG6QFC

A. Data

The dataset I chose to work with includes data on material and fetal outcomes in hospitals and other health facilities in Kenya and Uganda. The data is comprised of individual data points such as sex, discharge status, and Apgar scores for infants/fetuses, and age, discharge status, gestation, and type of delivery for mother. The data was derived from analyzing the maternity registers, records kept on all births in a health facility, from 23 different facilities in Kenya and Uganda, from the years 2016 through 2018.

There are a number of stakeholders for the data that I was able to identify through examining the data itself, and the accompanying publication. The 23 healthcare facilities surveyed are stakeholders of the data as they are ultimately the creators of the data. (Waiswa *et al.*, 2020) Four individuals, Brennan V. Higgins, Elizabeth Butrick, Rikita Merai, and Nancy L. Sloan, are identified as curators of the data, and are also stakeholders. (Waiswa *et al.*, 2020) The users of the data are also stakeholders, and they include the author of the accompanying journal article, as well as the additional researchers and contributors to the study and article. (Waiswa *et al.*, 2020) Other stakeholders of the data, who serve as interested parties to the research include the Preterm Birth Initiative Kenya & Uganda Implementation Research Collaborative, and the Kenyan and Ugandan Ministries of Health. (Waiswa *et al.*, 2020) Finally, the funder of the data is the Bill and Melinda Gates Foundation. (Waiswa *et al.*, Dataset, 2020)

The dataset is comprised of one data file, an .xlsx file, which contains 61,018 unique entries encompassing 22 unique data points or variables. In order to open the data file, you must have the ability to open an .xlsx file, which can be accomplished using several different software programs. The data is licensed under a CC0 1.0 (CC0 1.0) Public Domain Dedication license. This license means that there is no copyright on the work, and it can be copied, modified, and distributed without seeking permission. (*Creative Commons License Deed*)

B. Metadata

The data comes with two forms of metadata. The first is general information on the dataset, including methods of collections and contact information for the primary researchers. The metadata also includes a data dictionary to explain the variables used in the analysis. The metadata is located separately from the data files. It is included in two documents. The general information is included in a .txt file, which is easily readable using a number of different software programs. The data dictionary is uploaded as an .xlsx file, which is the same as the data itself. The methodological information included in the metadata is fairly comprehensive. It explains the process of data collection and analysis, including the software used to perform the statistical analyses. Additionally, there is full contact information for the primary researchers and data curator, which makes asking clarifying or specific questions about the data easier (so long as the contact information remains current). However, the metadata included relating to the variables themselves, including the data dictionary, is less comprehensive and somewhat scattered. It includes detailed descriptions of some variables, but other variables have no description, despite the use of uncommon acronyms in the variables themselves. One can use context clues to ascertain the meaning of those variables, but you run the risk of guessing wrong and drawing inaccurate conclusions based on the data. It would be ideal for that information to be included in the metadata along with the descriptions of the other variables.

The metadata does not appear to be structured according to any metadata standard that is known to this author. However, the .txt file does appear to be following, at least, an institution-specific template. The metadata included within the repository itself is quite accurate and robust to assist with discovery of the dataset. The dataset has been tagged with multiple, relevant keyword tags, and the natural language used in the abstract include multiple keywords that a researcher interested in these issues would use to search. I do feel that the data dictionary should be improved to ensure that all data elements are fully explained, as set forth above, so that an unfamiliar user could pick up this data and put it to use for their own purpose without needing to contact the researchers for additional information.

C. Publications

The dataset is associated with one published article, which is linked from the dataset’s repository profile. In order to determine if any other publications have used the dataset, I searched Google for the DOI citation for the dataset, as well as the title of the dataset. The results included in Google were for the dataset itself, or for the previously identified journal article. I also performed the same searches through the DU library and was only able to retrieve the previously identified article.

II. Data Repository Profile

Repository Selected: Population Services International Dataverse

Link to Repository: <https://dataverse.harvard.edu/dataverse/PSI>

A. Repository Policies and Procedures

I chose this repository because it has a concrete, and narrowly-tailored, purpose that speaks directly to my selected dataset. The repository’s parent organization is Population Services International (PSI) and its mission is stated as follows: “PSI is a global health organization dedicated to improving the health of people in the developing world by focusing on serious challenges like a lack of family planning, HIV and AIDS, barriers to maternal health.” (*Population services International (PSI) Dataverse*) As the dataset involves maternal and child mortality and birth outcomes in African countries, I feel this repository is an excellent fit. Additionally, the parent organization further states that its purpose is to work with local stakeholders to “create health solutions that are built to last.” (*Population services International (PSI) Dataverse*) This is an important feature for me, as I feel an important goal for sharing data should be for action and policymaking, especially in the public health realm. It is reassuring to know that the data will be used to inform on policies, instead of sitting dormant in the repository on the off chance someone else seeks to use the data in the future.

It was difficult to ascertain whether the repository is open for submissions based on the information associated with the repository. The website itself includes a dropdown to add datasets, which does require that you create an account with the Harvard dataverse. These accounts are limited to recognized institutions. (*See* *Log In - Harvard Dataverse*) However, it is unclear at that stage whether you are simply creating an account for storage within Harvard’s dataverse overall, or rather submitting a dataset directly to the PSI repository. However, in the absence of published policies, I examined the datasets available within the repository. Based upon the metadata it appears that the producer of each dataset is a subsidiary or affiliate of PSI. As such, I concluded that the data repository is not open for general submissions. (*See e.g.* Kaula & Buyungo, 2017)

Additionally, based upon the review of the metadata, I believe the collection scope of the repository is limited to researchers working for PSI (whether directly or by partnership) in one of the identified PSI coverage areas (Asia and Eastern Europe, East Africa, Latin America and the Caribbean, Southern Africa, and West and Central Africa). It does appear that submission is open to work funded by grants outside of PSI, however; the work must have been conducted by PSI affiliated researchers. There are no collection policies that are available on the website, which further leads me to believe that the repository is closed to general submissions.

B. Repository Data Access Mechanisms

Users do not need a login to download data from the repository. However, some of the data is embargoed or restricted because it contains potentially identifiable information (PII). There is a mechanism where non-PSI users may request access to restricted data files. Unfortunately, the link for the request form is currently a broken link, so I was unable to ascertain what additional information would be required to access the information. (*See* <https://www.psi.org/research/ethics-data-use-and-authorship/data-use-and-authorship/>) Additionally, some datasets require the user to fill out an entry in the data “guest book” indicating the user’s name, email address, and institution before access to the data is granted. (*See e.g.* Kaula & Buyungo, 2017)

The repository allows for direct file download and offers no other options for access. I suspect this is because of the access restrictions for some files. Metadata is displayed in a format which is proprietary to Dataverse, Dataverse 4.0. However, Dataverse 4.0 supports metadata schemas from DDI, DataCite, Dublin Core, Virtual Observatory, and ISA-Tab. (*See* Appendix) Dataverse also provides the user the option to export metadata in a variety of formats, including Dublin Core, JSON, OpenAire, and DataCite. (*See* *e.g.* Shaw et al., 2017)

The repository is structured so that each dataset stored within has its own landing page. From the landing page, the user can then access the file downloads for the data, as well as export the metadata. The DIP for each dataset varies in the sense that not every associated data file is available to the public. However, in reviewing multiple datasets there is a standard file format (a combination of .pdf, .xlsx, and .dta) and there are standard metadata included for each dataset (citation, geospatial, and PSI metadata). (*See* *e.g.* Shaw et al., 2017)

C. Conclusion

Overall, I had a difficult time finding a repository that was open for submission from the public with a collection development policy that fit the characteristics of my dataset. If I did find “open” repositories that were suitable for a public health/health focused dataset, upon further review they were limited to submission of datasets that accompanied a publication. It seems unusual that there are not many open data repositories for public health data. However, I can speculate that part of the issue has to do with the sensitive nature of health data, as it often contains identifiable information.

III. Final Information

A. Data Citation

My recommended data citation for this dataset is as follows:

Waiswa P, Higgins BV, Mubiri P, et al. (2021). Maternity Register Dataset. Harvard Dataverse. DOI############.

I formulated this data citation based on the Data Cite recommended format for data citation, (Team, D. C.) and used a dummy DOI as the identifier instead of the Github URL for my repository. I made this choice because all uploads to my repository are assigned a DOI, and because DOIs are a preferable method of identification to URLs.

B. Considerations for long-term preservation

At current, the files comprising this dataset are in a relatively stable, easily accessible file format (.xlsx and .txt). There is always the concern that Microsoft could choose to stop supporting Excel as a software program, however, that seems unlikely as this point. What is a more realistic concern is that as file formats and functionality evolve, the .xlsx format could become obsolete. As it is, the Harvard Dataverse supports .xlsx files, and does not support the predecessor .xls files. (*Dataset + File Management*) As the curator of this dataset, one will want to be mindful of this possibility and ensure that any evolutions in file formats will not cause the files stored within the repository to become obsolete or non-functional.

C. Copyright

All datasets uploaded into the Harvard Dataverse (including the PSI Dataverse) are automatically given a Creative Commons CC0 Public Domain License. (*Dataset + File Management*) This license allows for individuals to access and utilize the dataset without permission from the owner. (*Creative Commons License Deed*) Users can request that a different license be applied to their dataset; however, I believe that this license is appropriate based on the fact that the research was designed to inform health practices in developing countries.

D. Human Subject Considerations

This dataset is directly related to human subjects, as it is based on maternity register information from hospitals. While the data does contain potentially identifiable information, the facilities, countries, and patients have been deidentified. Additionally, while the data was collected over a two-year period, there are no dates associated with the individual records. Based on this, I do feel that this data is appropriate for free and open public use. However, one benefit of the Harvard Dataverse interface is that it allows for access to files within a repository to be restricted. So, should a concern arise regarding the possible identification of human subjects within the data, the access to the file could be restricted while still being maintained and preserved.

References

*Appendix*. https://guides.dataverse.org/en/5.3/user/appendix.html#user-appendix.

*Creative Commons License Deed*. Creative Commons - CC0 1.0 Universal. https://creativecommons.org/publicdomain/zero/1.0/.

*Dataset + File Management*. Dataverse.org. https://guides.dataverse.org/en/5.3/user/dataset-management.html.

Kaula, H., & Buyungo, P. (2017, March 3). *Uganda (2015): Practices and Perceptions about cervical cancer screening and preventive treatment among women of reproductive age (WRA) in Uganda. Round 1.* Harvard Dataverse. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F29616>.

Log In - Harvard Dataverse. https://dataverse.harvard.edu/loginpage.xhtml?redirectPage=%2Fdataset.xhtml%3FownerId%3D1.

*Population services International (PSI) Dataverse*. https://dataverse.harvard.edu/dataverse/PSI.

Shaw, B., Onema, W., Mpanga, G., Kasongo, G., Thanel, K., Riley, C., & Garfinkel, D. (2017, June 14). *DRC (2015): FPwatch Family Planning Outlet Survey*. Harvard Dataverse. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FOJD10N>.

Team, D. C. *Cite Your Data*. DataCite. https://datacite.org/cite-your-data.html.

Waiswa, Peter *et al.* (2020). Pregnancy outcomes in facility deliveries in Kenya and Uganda: A large cross-sectional analysis of maternity registers illuminating opportunities for mortality prevention, Dryad, Dataset, https://doi.org/10.7272/Q6ZG6QFC

Waiswa, P., *et al.* (2020). Pregnancy outcomes in facility deliveries in Kenya and Uganda: A large cross-sectional analysis of maternity registers illuminating opportunities for mortality prevention. *PLOS One*. https://doi.org/https://doi.org/10.1371/journal.pone.0233845