Statistics Corner: FAIR Data Sharing

Kamal Kishore¹, Rakesh Kapoor²

REALITY CHECK

Data is the foundation of research and vice versa. The replication and verification of study findings are hallmarks of scientific research. Therefore, there is increasing thrust by the scientific fraternity to share data openly. The sharing of data is one problem. However, preparing shareable data without compromising study characteristics is another. Therefore, data sharing is challenging and not widespread despite increasing thrust. In the context of data sharing, the investigator needs to know:

- What are the various phases in a typical data cycle?
- What are the challenges in each phase of the data cycle?
- What are the requirements of the data anonymization process?
- What are the consequences of the Personal Data Protection Bill (PDP Bill, 2019) on the researcher?

Keywords: Data anonymization, Data cycle, Data sharing, De-identification, FAIR principle.

Journal of Postgraduate Medicine Education and Research (2020): 10.5005/jp-journals-10028-1351

INTRODUCTION

A significant hallmark of health and clinical sciences is to understand the disease characteristics by collecting data. Primarily, randomized control trials (RCTs), cohort, case-control and cross-sectional study designs are used to study disease characteristics. Randomized control trials address known and unknown confounding issues in the experiment. These also require the highest standard of conduct with utmost caution. The regulatory bodies and investigators also regulate and monitor RCTs closely. Therefore, RCTs are considered gold standards to study drug efficacy and adverse reactions in the patients. The International Committee of Medical Journal Editors (ICMJE) made it mandatory to share data along with manuscripts for patients recruited in the RCTs from 1st January 2019 onwards. The investigators can select from different data share model proposed by ICMJE.¹

Any alteration in data share practices, data collection method, and inclusion–exclusion criteria after registration requires justifiable reasons for modifications at later stages. However, data generation and data sharing for study designs other than RCTs are not appropriately regulated. Still, there is growing clamor among research fraternity to share data for studies other than RCTs to address the reproducibility crisis in science. The onus of ethically collecting and sharing data lies with investigators. However, many researchers do not understand the nitty-gritty of preparing a standard data file for sharing. It is crucial to understand that data undergoes many phases from planning to publication and each phase has its challenges. Figure 1 displays the various phases of the data life cycle from planning to discarding. The data is shared multiple times among multiple stakeholders during different phases of the data cycle. Therefore, many investigators knowingly or unknowingly share raw data without anonymization.

The free flow of data without anonymization poses a significant threat to patient safety and privacy. Moreover, Govt. of India (GOI) has drafted the Personal Data Protection Bill (PDP Bill, 2019) to protect the privacy of individuals.² This bill applies to data processed at the individual level. The investigators of health sciences need to be extra careful due to the sensitive nature of data. The data sharing with professionals and publication houses

^{1,2}Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

Corresponding Author: Kamal Kishore, Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India, Phone: +91 9591349768, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Kapoor R. Statistics Corner: FAIR Data Sharing. *J Postgrad Med Edu Res* 2020;54(1):17–19.

Source of support: Nil

Conflict of interest: None

(such as statisticians, journal clubs, publication houses) outside the study aggravates the challenge further. However, sharing is essential from the perspective of increased transparency and replicability of the study results. Therefore, data sharing is a good practice and a welcome step. This article is an attempt to familiarize the investigators with the general process of data anonymization. Investigators need to carefully understand and review the process of data de-identification in general and in particular to their datasets, as each dataset is unique.

Fig. 1: Routine data cycle

Fig. 2: A list of identifier variables under different domains for anonymization of data

DATA PHASE CHALLENGES

The different phases of data can be broadly segmented into data management and data sharing challenges. Data management for large projects is difficult and time-consuming. There is numerous specialized data management software such as PostgreSQL, MongoDB, Oracle database and MySQL, which are resource-intensive in terms of time and expertise. However, the majority of people use spreadsheets and do not deal with large datasets. Therefore, this article is for general users of spreadsheet who want to anonymize and standardize the data for sharing and analysis. Initial hurdles can vary from the type, format, number of variables, naming of dataset and subsets and selection of appropriate software to capture and manage data. The next challenge is to code, clean and prepare data dictionary appropriately. It is essential to keep a separate copy of crude (unclean and uncoded) data as a backup for future references. More details can be found in the "Data Cleaning-I" article written under the biostatistics series.³ The sensitive nature of health data and right to privacy are also prominent parameters before sharing and discarding the data during and after the project. Therefore, investigators should carefully decide to store the data in a separate folder, separate computer or cloud-based platform with password protection.

DATA ANONYMIZATION

Data are the building blocks of scientific research. Data sharing is an essential component of the progress and replication of scientific research. Many publications stress the importance of data sharing.^{4,5} Publicly sharing of data also increases citation.⁶ It is essential to make a delicate balance between patient privacy and data utility. The researchers need to follow findability, accessibility, interoperability, and reusability (FAIR) principles for publication, discovery and, use of data.⁷ The process of data anonymization in this article is explained with an assumption that data is anonymized and subsequently used for genuine scientific research. The data can be broadly segmented into identifier and non-identifier variables for anonymization. The identifiers can be further segregated into four major domains. The suggestions and recommendations in the four major domains are general without being exhaustive. Figure 2 depicts the list of identifier variables under various domains. The identifiers singularly or in combination may lead to the identification of patients' characteristics. Initially, investigators need to generate a variable with random numbers for study participant in the original or raw spreadsheet. Subsequently, make a copy of the raw spreadsheet and delete identifiers from the sheet. Afterward, follow the principle of data cleaning and coding explained in previous articles of the series.³ However, investigators need to be cautious

and carefully consider the identifiers while anonymizing data. Finally, use various platforms such as the Dataverse Project (<https://dataverse.org/>), DataCite (<https://datacite.org/>), Mendeley (<https://data.mendeley.com/>), DataHub (<https://datahub.io/>), Figshare (<https://figshare.com/>) for sharing of data. Nature publishing also keeps a record of various data repository platforms (<https://www.nature.com/sdata/policies/repositories>).

DATA PROTECTION RULE

Data sharing facilitates research. However, the sharing of raw data without anonymization is unethical and punitive by law. Many scientists share data in raw and summarized forms at various stages (conferences, journal clubs, data analyst and journal publication houses). Many times, adequate precautions are not taken to preserve the identity of study participants. The UK Data Service highlights the five safes: "Safe Projects", "Safe Settings", "Safe People", "Safe Outputs" and "Safe Data" to integrate into "Safe Sharing" of data.⁸ The right to privacy is a fundamental right granted by the constitution of India. Therefore, researchers strictly need to follow data anonymization practices in routine. The reasons to share without de-identifying may vary from the dearth of guidelines to no familiarity with the process of de-identification. The PDP Bill 2019 applies to the collection, disclosure, and processing of personal data collected within India. The act mandates the data collector and processor to take necessary steps to protect, de-identify, prevent misuse and unauthorized access. It also states that the sharing of sensitive data for processing with outsiders needs to be through a valid contract. Similarly, cross-border transfer of data to be carried out as per contract specified in the act. There is also a provision of imprisonment and fine for failing to meet the standard laid down in the act.

CONCLUSION

There are no standard guidelines for data sharing and anonymization except RCTs. Therefore, data anonymization and sharing are not standard practices among the majority of the investigators. Knowledge of the same is essential since investigators share data with various stakeholders outside the study. We have enlisted various variables of importance for data anonymization. These variables should be carefully screened and used in regular practice. The regular practice of data anonymization will help researchers refine the skills for de-identification.

ACKNOWLEDGMENTS

The authors would like to thank Dr Nipun Verma and Dr Sabina Regmi for taking time out of their busy schedule to review this article.

REFERENCES

1. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the international committee of medical journal editors. *JAMA* [Internet] 2017;317(24):2491–2492. DOI: 10.1001/jama.2017.6514.
2. Govt. of India. The personal data protection bill; 2019.
3. Kishore K, Kapoor R, Singh A. Statistics corner: Data cleaning-I. *J Postgrad Med Educ Res* 2019;53(3):130–132. DOI: 10.5005/jp-journals-10028-1330.
4. Piwowar HA, Becich MJ, Bilofsky H, et al. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med* 2008;5(9):1315–1319. DOI: 10.1371/journal.pmed.0050183.
5. Federer LM, Lu YL, Joubert DJ, et al. Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. *PLoS One* 2015;10(6):1–17. DOI: 10.1371/journal.pone.0129506.
6. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. *PLoS One* 2007;2(3):e308. DOI: 10.1371/journal.pone.0000308.
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:1–9. DOI: 10.1007/s40745-015-0066-4.
8. <http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/>. UK Data Service. Access to sensitive data for research: 'The 5 Safes' [Internet]. 2015 [cited 2020 Jan 20]. Available from: <http://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/>.