COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Implementing Informatics Tools with Data Management Plans for Disease Area Research

COLLECTION:
DATA MANAGEMENT PLANNING ACROSS DISCIPLINES AND INFRASTRUCTURES

PRACTICE PAPER

VIVEK NAVALE
MATTHEW MCAULIFFE

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Data Management Plans (DMPs) are essential to a research data life cycle. The DMPs should be developed as part of the research programs to be effective. For disease area research, integrating research community-recommended data standards during collection can enhance the likelihood of data reuse. Informatics tools are required as part of DMPs with the aim of data being findable, accessible, interoperable, and reusable.

The US National Institutes of Health supports various disease area research programs and has recently finalized the Data Management and Sharing Policy. The policy highlights the importance of sharing data and metadata, including information on various elements such as data types, standards, storage repositories, access, services, and tools used for a proposed research project.

The present paper provides Traumatic Brain Injury (TBI) and Parkinson's Disease (PD) research as examples of where the elements of the policy are being supported. The software tools that have been developed for the TBI and PD plans are available through the Biomedical Research Informatics Computing System. A Protocol and Form Research Management System (ProFoRMS) facilitates researchers to manage research protocols when collecting clinical data. The ProFoRMS also supports automatic validation with the data dictionaries for TBI and Parkinson's disease. Detailed information on the functionality of the software tools used for preserving data within TBI and PD repositories is openly available on their respective websites.

CORRESPONDING AUTHOR:
Dr. Vivek Navale
Center for Information Technology, 9000 Rockville Pike, Bldg. 12A, (Rm 4041), National Institutes of Health, Bethesda, Maryland, 20892, USA
Vivek.Navale@nih.gov

KEYWORDS:
Data Management Plan; Informatics Tools; Traumatic Brain Injury; Parkinson's Disease

TO CITE THIS ARTICLE:
Navale, V and McAuliffe, M. 2023. Implementing Informatics Tools with Data Management Plans for Disease Area Research. *Data Science Journal*, 22: 24, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2023-024

INTRODUCTION

Data Management Plans (DMPs) serve as planning tools to describe the type of data and metadata produced during a research project, the standards used during the collection of data, and the designated repositories for storage, along with information on associated software tools required for accessibility. Traditionally, DMPs have been considered static documents, human-readable in nature; however, more recently, there is an impetus for DMPs to be machine-actionable. Embedding DMPs in existing workflows can improve the stakeholder experience, involving funding agencies, researchers, repository managers, publishers, and others (Miksa et al. 2019). DMPs are integral in a research data life cycle, which involves the collection, processing, validation, storage, analysis, and access (Williams, Bagwell & Nahm Zozus 2017). Recently, the US National Institutes of Health finalized the policy for Data Management and Sharing (NOT-OD-21-013: Final NIH Policy for Data Management and Sharing). The policy highlights the importance of good data management practices and enables sharing of scientific data generated from NIH-funded or conducted research. The policy is supported by supplemental information, including privacy protection when sharing Human Research Participant Data and the elements for Data Management and Sharing Plans.

In biomedical clinical research, typically, research hypotheses are tested, and the results depend on the accuracy and reliability of data collected during the study. Clinical Data Management involves various stages that can be integrated by implementing DMPs. The plans comprise designing Case Report Forms (CRFs) to organize clinical protocol-specific information collected for a research project. The data fields in CRFs should be clearly defined and consistent throughout. It should be supported by data and metadata submission procedure specifications, validation and discrepancy resolution methodology, extraction, coding, audit trail management, and secure storage within a database (Krishnankutty et al. 2012).

The CRFs traditionally have been paper-based, involving hand-written notes, however, electronic CRFs are increasingly being used to provide consistency during the clinical research data collection work. The CRF development work occurs at the very beginning of the research study, during the protocol development and approval process. The responsibilities of the CRFs primarily reside with the Principal Investigator and the team conducting the research study. As a best practice for clinical research, only required data should be collected. By using standards that apply to the research protocols, the CRFs can further increase data quality during a research study (Gazali, Kaur & Singh 2017).

Informatics software can be used to support clinical DMPs. It can facilitate the creation of CRFs, import existing paper-based forms into electronic CRFs, and provide procedures for validating the data collected during the study (Brandt et al. 2006). Recently, clinical DMPs with mobile applications have been used for data collected for a community-level disease study (Oviedo Sarmiento et al. 2021).

In this paper, we discuss informatics tools that researchers utilize to support clinical DMPs for advancing Traumatic Brain Injury (TBI) and Parkinson's Disease (PD). The tools are part of a Biomedical Research Informatics System (BRICS) developed for several National Institutes of Health (NIH) biomedical research programs (Navale et al. 2019). We discuss the software services that are used in the research projects and explain the association of the informatics tools for supporting the NIH Data Management and Sharing plan elements. The tools discussed lead to sharing of de-identified data with the goal of data being FAIR (Wilkinson et al. 2016), (findable, accessible, interoperable, and reusable) within designated disease-specific digital repositories that are trustworthy (Lin et al. 2020).

DATA COLLECTION

The NIH and the US Department of Defense provide research grants that support TBI and PD studies and require submission of data to the designated data platforms (*FITBIR: Federal Interagency Traumatic Brain Injury Research Informatics System*), and Parkinson's Disease Biomarker Program (*PDBP*), respectively. Within a disease area, (e.g., TBI and Parkinson's) clinical research is protocol-specific, which involves data collection on specific variables (data elements) that enable the data analyses to prove or disprove research hypotheses. For TBI and Parkinson's researchers, using Common Data Elements (CDEs) as part of the clinical DMP is recommended (Grinnon et al. 2012). A CDE is defined as a fixed representation of a variable collected within a clinical domain, interpretable unambiguously in human and machine-readable forms (Navale et al. 2018). The National Institute of Neurological Disorders and Stroke (NINDS) provides detailed information on CDEs and has also developed CRF templates *(CRF library)* with data dictionaries for different neurological diseases**.** The data dictionaries comprise data elements, form structures, and electronic forms. A data element has a name and precise definition with permissible values when applicable. A data element directly relates to a question on a form, and the form structure serves as the container for data elements.

As a first step, researchers are asked to complete a data submission form (a component of DMP) that is reviewed and approved by the data access committee. Creating a research study in the designated data repository (e.g., TBI and PD repositories) and associating the submission form is part of the data management process. The study metadata includes information on the organization, Principal Investigator, funding source and IDs, study type(s), start and end dates for grants, therapeutic agents used, sample size, data types, forms used, and publications.

A Protocol and Form Research Management System (ProFoRMS) provides researchers with the services to manage research protocols when collecting clinical data. An electronic CRF system enables scheduling patient visits, collecting, adding new data, modifying previously collected data entries, and correcting any discrepancies before submission to designated repositories. The software tools used to execute the ProFoRMS have been discussed in an earlier publication (Navale & McAuliffe 2022). The CDE-based data dictionaries are available with the Federal Traumatic Brain Injury and Parkinson's Disease Biomarker (PDBP) data platforms. The ProFoRMS also supports automatic validation with the data dictionaries for TBI and Parkinson's disease. Researchers have the option to collect data by other methods (e.g., REDCap), however, the output files from other methods are validated with the data dictionaries before submitting to the designated repositories for TBI and Parkinson's disease.

DATA DE-IDENTIFICATION

A random alphanumeric unique identifier that is not directly generated from personally identifiable information (PII) is assigned to individual patients. The BRICS privacy-preserving record linkage tool, also known as the Global Unique Identifier (GUID), creates one-way encrypted hash codes, allowing the PII to reside only on the researcher's site. The GUID tool (Johnson et al. 2010) is available through the FITBIR and PDBP platforms. This approach of using unique identifiers allows for the tracking of patients who may be enrolled in multiple studies (Navale et al. 2018).

DATA VALIDATION

The file format for data submissions is comma-separated values and is structured to be consistent with CDE-variable names and data values. A validation tool is available to support the data repositories and ProFoRMs modules. The tool compares the submitted values with the defined and/or acceptable ranges for TBI and Parkinson's disease CDEs. Any identified errors during this process are corrected before a data submission package is produced for uploading to a designated repository.

DATA STORAGE

Research study information is stored within the TBI and Parkinson's disease repositories. Management of a research study within a repository prompts researchers to describe in detail the data collected to make data accessible to users. The information within repositories contains patient assessment (form) data, imaging, electroencephalogram, magnetoencephalography, and derived genomics data.

DATA ACCESS

As raw data is stored in the repositories, initially, access is limited to the Principal Investigator and their team members who can share with other researchers associated with their work. Currently, the TBI and PD data access committee permits researchers to maintain the data in a private state for a year after the research grant has ended, but after one year, the data moves to a shared state. All approved data users have access to the shared data. The data repository also provides an interface for generating digital object identifiers for a study that can be referenced in research articles**.**

Metadata and study summaries are also made available via FITBIR and PDBP public sites. The FITBIR provides a metadata visualization tool that helps in searching for research studies (https://fitbir. nih.gov/visualization).

Table 1 illustrates the various elements of the NIH Data Management and Sharing Plan. The plan highlights the importance of providing required information on the amount and type of data (e.g., imaging, genomic, mobile, survey) being collected, the level of aggregation (e.g., individual, aggregated, summarized), and the degree of data processing that has occurred (i.e., raw or processed data). It requires information on standards (data formats, dictionaries, identifiers, definitions, and associated documentation) used when collecting data. Also, the plan should provide information on how data and metadata will be findable and identifiable (e.g., persistent unique identifier or other indexing tools), maintain privacy and confidentiality (i.e., de-identification, certificates of confidentiality, and other protective measures), and identify the repository(ies) where the scientific data and metadata will be preserved. Information should also indicate when the scientific data and metadata will be made available to other users and specify the duration of accessibility to the data.

Table 1 Associating Data Management and Sharing Plan Elements with Informatics Tools.

DATA TYPE	DATA STANDARD	DATA PRESERVATION	DATA ACCESS AND SHARING	SERVICES AND TOOLS	OVERSIGHT OF DATA MANAGEMENT
TBI Patient clinical data, imaging data, bio-samples	FITBIR Common Data Elements, Data dictionaries	FITBIR repository, CSV files, DICOM images	Controlled access, DAC approvals needed, Meta-studies	BRICS service modules – GUID, ProFoRMS, FITBIR Repository, Query tool	DOD, CIT, and NINDS https://fitbir.nih.gov/
PD Patient clinical data, imaging data, bio-samples, genomics data (VCF)	PDBP Common Data Elements, Data Dictionaries	PDBP repository, CSV files, DICOM images	Controlled access, DAC approvals needed, Meta-studies	BRICS service modules – GUID, ProFoRMS, PDBP repository, Query tool	CIT and NINDS https://pdbp.ninds.nih.gov/
Eye disease clinical data, genomics data (processed)	NEI Common Data Elements, Data Dictionaries, LOINC data standards	NEI repository, CSV files,	Controlled access, DAC approvals needed, Meta-studies	BRICS service modules- GUID, ProFoRMS, NEI repository, Query tool	NEI and CIT https://eyegene.nih.gov

In light of the NIH Data Management and Sharing Policy, TBI, PD, and eye disease are provided as examples to illustrate the information needed for the plan elements. The type of data collected, the CDEs used during collection, and the designated repositories for each of the examples are unique to biomedical programs. For the examples shown in Table 1, the services and software tools used are similar.

FAIR DATA AND TRUSTWORTHY REPOSITORIES

Clinical DMPs play an important role in enabling data to be FAIR. Maintaining the confidentiality of the patients is of paramount importance. The DMPs for TBI and PD require data de-identification as a prerequisite step during data processing. With the assignment of unique identifiers, data from the same patient is findable and can be integrated as needed. Data accessibility is governed by the data repository policies and requires approval by the data access committees. The metadata summaries are publicly available through the FITBIR and PD data platforms.

The use of data dictionaries and associated CDEs for TBI and PD research studies provides consistency in data collection, improves data quality, and facilitates integrating data for different studies during analysis work. Also, the adoption of research community-recommended standards during data collection and subsequent preservation in centralized repositories for TBI and PD leads to the trustworthiness of data.

CONCLUSION

Clinical DMPs are an integral part of the research data life cycle process that involves collection, processing, validation, and storage within repositories. Informatics tools provide direct support for clinical DMPs to effectively establish confidentiality, integrity, and accuracy of clinical records. Maintaining patient data confidentiality is essential within a clinical setting. To ensure data quality during the de-identification process, it is important to utilize clearly defined data concepts/variables, data dictionaries, and systems that support data curation and preservation (AbuHalimeh 2022). Electronic data collection with eCRFs that are associated with data dictionaries can reduce the time involved in data curation work. Validation of data before storage improves the data quality and promotes trustworthiness in repositories.

ABBREVIATIONS

BRICS – Biomedical Research Informatics Computing System,

DAC – Data Access Committee,

DICOM – Digital Imaging and Communications in Medicine,

GUID – Global Unique Identifier,

ProFoRMS – Protocol Form and Research Management System,

TBI – Traumatic Brain Injury,

FITBIR – Federal Interagency Traumatic Brain Injury Research,

DOD -Department of Defense,

CIT – Center for Information Technology,

NINDS – National Institute of Neurological Disorders and Stroke,

PDBP – Parkinson's Disease Biomarker Program,

LOINC – Logical Observation Identifiers Names and Codes,

NEI – National Eye Institute,

VCF – Variant Call Format.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Vivek Navale orcid.org/0000-0002-7110-8946
Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA
Matthew McAuliffe orcid.org/0000-0002-2409-5126
Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA

REFERENCES

AbuHalimeh, A. 2022. Improving Data Quality in Clinical Research Informatics Tools. *Frontiers in Big Data*, 5: 1–6. DOI: https://doi.org/10.3389/fdata.2022.871897

Brandt, CA, et al. 2006. Informatics tools to improve clinical research study implementation. *Contemporary Clinical Trials*, 27(2): 112–122. DOI: https://doi.org/10.1016/j.cct.2005.11.013

Case Report Form (*CRF library*). Available at: https://commondataelements.ninds.nih.gov/crf-library (Accessed: 13 October 2022).

***Federal Interagency Traumatic Brain Injury Research Informatics System* (FITBIR).** Available at: https://fitbir.nih.gov/ (Accessed: 6 November 2017).

Gazali, KS and **Singh, I.** 2017. Artificial intelligence based clinical data management systems: A review. *Informatics in Medicine Unlocked*, 9: 219–229. DOI: https://doi.org/10.1016/j.imu.2017.09.003

Grinnon, ST, et al. 2012. National Institute of Neurological Disorders and Stroke Common Data Element Project – approach and methods. *Clinical Trials*, 9(3): 322–329. DOI: https://doi.org/10.1177/1740774512438980

Johnson, SB, et al. 2010. Using global unique identifiers to link autism collections. *Journal of the American Medical Informatics Association: JAMIA*, 17(6): 689–695. DOI: https://doi.org/10.1136/jamia.2009.002063

Krishnankutty, B, et al. 2012. Data management in clinical research: An overview. *Indian Journal of Pharmacology*, 44(2): 168–172. DOI: https://doi.org/10.4103/0253-7613.93842

Lin, D, et al. 2020. The TRUST Principles for digital repositories. *Scientific Data*, 7(1): 1–5. DOI: https://doi.org/10.1038/s41597-020-0486-7

Miksa, T, et al. 2019. Ten principles for machine-actionable data management plans. *PLoS Computational Biology*, 15(3): 1–15. DOI: https://doi.org/10.1371/journal.pcbi.1006750

Navale, V, et al. 2018. Standardized Informatics Computing Platform for Advancing Biomedical Discovery Through Data Sharing. *bioRxiv*. DOI: https://doi.org/10.1101/259465

Navale, V, et al. 2019. Development of an informatics system for accelerating biomedical research. *F1000Research*, 8: 1–19: DOI: https://doi.org/10.12688/f1000research.19161.2

Navale, V and **McAuliffe, M.** 2022. The Integration of a Canonical Workflow Framework with an Informatics System for Disease Area Research. *Data Intelligence*, 4(2): 186–195: DOI: https://doi.org/10.1162/dint_a_00125

NOT-OD-21-013: Final NIH Policy for Data Management and Sharing. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (Accessed: 2 June 2023).

Oviedo Sarmiento, OJ, et al. 2021. Data management plan for a community-level study of the hidden burden of cutaneous leishmaniasis in Colombia. *BMC Research Notes*, 14(1): 1–6: DOI: https://doi.org/10.1186/s13104-021-05618-4

Parkinson's Disease Biomarker Program (PDBP). Available at: https://pdbp.ninds.nih.gov/how-to-guide (Accessed: 20 December 2018).

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 1–9: DOI: https://doi.org/10.1038/sdata.2016.18

Williams, M, Bagwell, J and **Nahm Zozus, M.** 2017. Data management plans: the missing perspective. *Journal of Biomedical Informatics*, 71: 130–142. DOI: https://doi.org/10.1016/j.jbi.2017.05.004

TO CITE THIS ARTICLE:
Navale, V and McAuliffe, M. 2023. Implementing Informatics Tools with Data Management Plans for Disease Area Research. *Data Science Journal*, 22: 24, pp. 1–6. DOI: https://doi.org/10.5334/dsj-2023-024

Submitted: 07 November 2022
Accepted: 22 June 2023
Published: 23 August 2023