Essentials of Data Management: An Overview

Miren B. Dhudasia, MBBS, MPH[1,2], **Robert W. Grundmeier, MD**[2,3,4], **Sagori Mukhopadhyay, MD, MMSc**[1,2,3]

[1]Division of Neonatology, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[2]Center for Pediatric Clinical Effectiveness, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[3]Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States

[4]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, United States

What is data management?

Data management is a multi-step process that involves obtaining, cleaning, and storing data to allow accurate analysis and produce meaningful results. While data management has broad applications (and meaning) across many fields and industries, in clinical research the term data management is frequently used in context of clinical trials.[1] This editorial is written to introduce early career researchers to practices of data management more generally, as applied to all types of clinical research studies.

Outlining a data management strategy prior to initiation of a research study plays an essential role in ensuring that both scientific integrity (i.e., data generated can accurately test the hypotheses proposed) and regulatory requirements are met. Data management can be divided into three steps – data collection, data cleaning and transformation, and data storage. These steps are not necessarily chronological and often occur simultaneously. Different aspects of the process may require expertise of different people necessitating a team effort for effective completion of all steps.

1. Data collection

Data source: Data collection is a critical first step in the data management process and may be broadly classified as "primary data collection" (collection of data directly from the subjects specifically for the study) and "secondary use of data" (re-purposing data that was collected for some other reason – either for clinical care in the subject's medical record or

Corresponding author: Sagori Mukhopadhyay, MD, MMSc. Roberts Center for Pediatric Research, Children's Hospital of Philadelphia, 2716 South Street, Office 19-322, Philadelphia, PA 19146. MUKHOPADHS@chop.edu. Phone: (215) 829-3603. Fax: (215) 829-7211.

Disclosure statement: The authors have no conflicts of interest or financial relationships relevant to this article to disclose.

for a different research study). While the terms retrospective and prospective data collection are occasionally used,[2] these terms are more applicable to how the data is *utilized* rather than how it is *collected*. Data used in a retrospective study is almost always secondary data; data collected as part of a prospective study typically involves primary data collection, but may also involve secondary use of data collected as part of ongoing routine clinical care for study subjects. Primary data collected for a specific study may be categorized as secondary data when used to investigate a new hypothesis, different from the question for which the data was originally collected. Primary data collection has the advantage of being specific to the study question, minimize missingness in key information, and provide opportunity for data correction in real-time. As a result, this type of data is considered more accurate but increases the time and cost of study procedures. Secondary use of data includes data abstracted from medical records, administrative data such as from hospital's data warehouse or insurance claims, and secondary use of primary data collected for a different research study. Secondary use of data offers access to large amounts of data that is already collected but often requires further cleaning and codification to align the data with the study question.

Data forms: A case report form (CRF) is a powerful tool for effective data collection. A CRF is a paper or electronic questionnaire designed to record pertinent information from study subjects as outlined in the study protocol.[3] CRFs are always required in primary data collection but can also be useful in secondary use of data to pre-emptively identify, define and if necessary, derive critical variables for the study question. For instance, medical records provide a wide array of information that may not be required or be useful for the study question. A CRF with well-defined variables and parameters helps the chart reviewer focus only on the relevant data, and makes data collection more objective and unbiased, and in addition, optimize patient confidentiality by minimizing the amount of patient information abstracted. Tools like REDCap (Research Electronic Data Capture) provide electronic CRFs and offer some advanced features like setting validation rules to minimize errors during data collection.[4] Designing an effective CRF upfront during the study planning phase helps to streamline the data collection process, and make it more efficient.[3]

2. Data cleaning and transformation

Quality checks: Data collected may have errors that arise from multiple sources - data manually entered in a CRF may have typographical errors, whereas data obtained from data warehouses or administrative databases may have missing data, implausible values, and non-random misclassification errors. Having a systematic approach to identify and rectify these errors, while maintaining a log of the steps performed in the process, can prevent many roadblocks during analysis.

First, it is important to check for missing data. Missing data is defined as values that are not available and that would be meaningful for analysis if they were observed.[5] Missing data can bias the results of the study depending on how much data is missing and what is the pattern of distribution of missing data in the study cohort. Many methods for handling missing data have been published. Kang et al provide a practical review of methods for handling missing data.[6] If missing data cannot be retrieved and is limited to only a small number of subjects, one approach is to exclude these subjects from the study. Missing data in different variables across many subjects often require more sophisticated approaches to account for the ‘missingness’. These may include creating a category of ‘missing’ (for categorical variables), simple imputation (e.g., substituting missing values in a variable with average of non-missing values in the variable), or multiple imputation (substituting missing values with the most probable value derived from other variables in the dataset).[7]

Second, errors in the data can be identified by running a series of data validation checks. Some examples for data validation rules for identifying implausible values are shown in Table. Automated algorithms for detection and correction of implausible values may be available for cleaning specific variables in large datasets (e.g. growth measurements).[8] After identification, data errors can either be corrected, if possible, or can be marked for deletion. Other approaches, similar to those for dealing with missing data, can also be used for managing data errors.

Data transformation: The data collected may not be in the form required for analysis. The process of data transformation includes re-categorization and re-codification of the data which has been collected along with derivation of new variables, to align with the study analytic plan. Examples include categorizing body mass index collected as a continuous variable into under- and over-weight categories, re-coding free text values such as “growth of an organism” or “no growth”, etc. into a binary ‘positive’ or ‘negative’, or deriving new variables such as average weight per year from multiple weight values over time available in the dataset. Maintaining a code-book of definitions for all variables, pre-defined and derived, can help a data analyst better understand the data.

3. Data storage

Securely storing data is especially important in clinical research as the data may contain protected health information (PHI) of the study subjects.[9] Most institutes that support clinical research have guidelines for safeguards to prevent accidental data breeches.

Data is collected in paper or electronic formats. Paper data should be stored in secure file cabinets inside a locked office at the site approved by the institutional review board. Electronic data should be stored on a secure approved institutional server of the site, and should never be transported using unencrypted portable media devices (e.g. “thumb drives”). If all study team members do not require access to study data, then selective access should be granted to the study team members based on their roles.

Another important aspect of data storage is data de-identification. Data de-identification is a process by which identifying characteristics of the study participants are removed from the data, in order to mitigate privacy risks to individuals.[10] Identifying characteristics of a study subject includes name, medical record number (MRN), date of birth/death, etc. To de-identify data, these characteristics should either be removed from the data or modified (for example, changing MRN to study IDs, changing dates to age/duration, etc.). If feasible, study data should be de-identified when storing. If you anticipate that re-identification of the study participants may be required in future then the data can be separated into two files, one containing only the de-identified data of the study participants, and one containing all the identifying information, with both files containing a common linking variable (e.g., study

ID), which is unique for every subject or record in the two files. The linking variable can be used to merge the two files when re-identification is required to carry out additional analyses or to get further data. The link key should be maintained in a secure institutional server accessible only to authorized individuals who need access to the identifiers.

To conclude, effective data management is important to the successful completion of research studies and to ensure the validity of the results. Outlining the steps of the data management process upfront will help streamline the process and reduce the time and effort subsequently required. Assigning team members responsible for specific steps and maintaining a log, with date/time stamp to document each action as it happens, whether you are collecting, cleaning or storing data, can ensure all required steps are done correctly and identify any errors easily. Effective documentation is a regulatory requirement for many clinical trials and is helpful for ensuring all team members are on the same page. When interpreting results, it will serve as an important tool to assess if the interpretations are valid and unbiased. Lastly, it will ensure the reproducibility of the study findings.

Acknowledgments

Statement of financial support: This work was partially supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health grant (K23HD088753).

References

1. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012;44(2):168–172. [PubMed: 22529469]
2. Weinger MB, Slagle J, Jain S, et al.Retrospective data collection and analytical techniques for patient safety studies. J Biomed Inform. 2003;36(1–2):106–119. [PubMed: 14552852]
3. Avey MCase Report Form Design. In: Rondel RK, Varley SA, Webb CF, eds. Clinical Data Management. 2nd ed. West Sussex, England: John Wiley & Sons Ltd. 1999:47–73.
4. Harris PA, Taylor R, Thielke R, et al.Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–381. [PubMed: 18929686]
5. Little RJ, D'Agostino R, Cohen ML, et al.The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355–1360. doi: 10.1056/NEJMsr1203730 [doi]. [PubMed: 23034025]
6. Kang HThe prevention and handling of the missing data. Korean journal of anesthesiology. 2013;64(5):402. [PubMed: 23741561]
7. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.
8. Daymont C, Ross ME, Russell Localio A, Fiks AG, Wasserman RC, Grundmeier RW. Automated identification of implausible values in growth data from pediatric electronic health records. Journal of the American Medical Informatics Association. 2017;24(6):1080–1087. [PubMed: 28453637]
9. Office for Civil Rights, Department of Health and Human Services. Health insurance portability and accountability act (HIPAA) privacy rule and the national instant criminal background check system (NICS). final rule. Fed Regist. 2016;81(3):382–396. [PubMed: 26742185]
10. (OCR), Office for Civil Rights. Methods for de-identification of PHI. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.Updated 2012. Accessed Nov 13, 2020.

Table.

Examples of validation rules that can assist in data quality checks.

Description	Example	Comment
Assessing for implausible value	Birth weight >7000 grams or <200 grams	Specifically, useful for continuous variables
Assessing for implausible relationships between two different variables	Gestational age at birth 40 weeks or more, but birth weight <1500 grams	This assessment can also identify outliers, where the relationship maybe correct but rare
Assessing for discrepancy in longitudinal relationship	Study tests obtained on subjects dated prior to date of birth	Calculating differences between chronological dates can quickly identify implausible dates
Assessing for expected distribution of common variables	Delivery rate at a site in one year much different from years prior to or after	May require data transformation and basic analysis to produce summary data