www.ajbrui.org

Afr. J. Biomed. Res. Vol. 23 (Supplement 2, December, 2020); 31- 33

Workshop proceeding

Data Management in Clinical Research

Akinyemi J.O.

Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan. Ibadan, Nigeria.

ABSTRACT

No matter the complexity of a clinical research, one critical component that serves as bridge between conceptualization of research idea and the eventual publication of findings is data management. It encapsulates the whole processes involved in managing the life cycle of data in a clinical research project. Data management can make or mar any research. For example, a research with scientifically sound design but poor, incoherent data management procedure would end up with poor quality data. In this article, the author provides a brief overview of key considerations for data management in clinical research.

Keywords: *Data planning; Data collection and storage; Data cleaning; Data sharing*

**Author for correspondence: Email: joakinyemi@com.ui.edu.ng, odunjoshua@gmail.com; Tel: +2348077677427*

Abstracted by:

Bioline International, African Journals online (AJOL), Index Copernicus, African Index Medicus (WHO), Excerpta medica (EMBASE), CAB Abstracts, SCOPUS, Global Health Abstracts, Asian Science Index, Index Veterinarius

INTRODUCTION

Data management covers the different tasks carried out under the following subheadings:

- Research data planning
- Data collection
- Data entry/storage
- Data cleaning
- Data analysis and interpretation
- Data sharing

Each of the processes itemized above are briefly described except data analysis and interpretation which requires the application of statistical methods. Comprehensive exposition on data analysis techniques and approaches are available in several other sources ((Peat and Barton, 2005; Kirkwood and Sterne, 2004; Daniel 2006).

RESEARCH DATA PLANNING

This is conveniently a part of research conceptualization and design. Research data planning is a wise investment for researches because at this stage, primary hypothesis, primary and secondary outcome measures are defined. The recommended good practices for data planning include the following:

- Review each variable or outcome to be measured for the best data type [nominal, ordinal or continuous]. Quantitative variables should be collected at the scale of measurement that gives the highest precision. For instance, it is better to collect "age at last birthday" instead of using age groups.
- Be exact and discrete when defining variables. Ensure that variable definitions are not subject to the whims of data collectors or interviewers.
- Code categorical variables during data collection. It is helpful to provide codes for close-ended questions. This makes the job of data entry clerks easier. It also makes it easy to design validated data entry processes thereby minimizing data entry errors.
- Be consistent, that is, similar questions and responses should use similar codes. For example, if Yes and No are coded as 1 and 2, this should be consistent throughout the data collection instrument or questionnaire.
- Case Report Form (CRF)should be designed along with protocol to assure collection of only those data the protocol specifies. For complex studies in which CRF move from one point to another, organization of the CRF should follow the same sequence in which research participants are seen or interacted with. Source documents for different variables should be clearly identified. Specify units of measurements; indicate how measurements are to be taken i.e. oral, rectal or axillary. Date format must be consistent throughout the CRF (avoid mixing both British and American date formats). Standard operating procedure manuals should be developed for all research staff. All data collection forms should be reviewed by all collaborators and study managers prior to study commencement.

DATA COLLECTION

Method of data collection will depend on structure or nature of data to be collected. This may be in paper form, electronic data capture, structured complex machine (Captured Data), biometric or physiological measurements (e,g. EEG, ECG), images (e.g. Radiological or Histological images). When data is to be collected from different service points (e.g. clinic, lab, radiology), it is necessary to provide a mechanism to ensure smooth linkage of data from different sources. A common practice is to assign study identification numbers at point of first contact. This ID is used on all study instruments for easy linkage. Depending on the study setting, data collection can be automated using computer-aided devices with which data are collected electronically and transmitted to central servers for synchronization. However, this requires careful planning and engagement of data management experts to ensure a seamless operation.

DATA ENTRY AND STORAGE

Technological advancement and proliferation of computers has made data entry and storage relatively easier than it was about two decades ago. A common practice now is to employ data clerks who will enter data using computer software. Some other researchers may also capture data electronically in which case, manual data entry by clerks becomes unnecessary. The use of computer software allows storage of large quantities of data, ease of checking and correction of errors, easy tabulation, presentation of results and quick statistical analysis.

The choice of data management software may be influenced by organizational preference, user-friendliness, portability, ease of data conversion, availability and system requirement. Common examples of data management software are REDCap, CSPro, Epi Info, Epi Data, SPSS, Ms Excel; Access etc. In some research institutes, customized software is developed for data management purposes.

Irrespective of the software used for data entry, it is advisable to have a data dictionary (or codebook) which defines the structure of the dataset and also enhance consistency. Occasionally, some investigators adopt "double entry" in which two independent data clerks enter the same data. The two datasets are subsequently compared to identify errors arising from inconsistencies which are resolved during data cleaning. The disadvantage of this approach is that it takes twice as long to enter the data, and it may have major cost or time implications.

DATA CLEANING

This is the process of detection and correction of errors in a dataset. Depending on the choice of data entry software, some data quality control measures can also be instituted at the time of data entry. This is common in software platform such as REDCap, CSPro, Epi Info and Epidata. Check codes are programmed for close-ended questions such that only valid entries are accepted. Inspite of quality control measures during data entry, errors may still be present. To provide good quality data for analysis, further cleaning is recommended. This can include checks for logical consistency across variables, range checks, missing data and exploratory data analysis. Sometimes, data cleaning may require that source documents be revisited. Depending on the software platform, a lot of data cleaning procedures can be automated by writing computer programs.

DATA SHARING

A new development in the research ecosystem is the concept of data sharing. In the past, investigators may share data with colleagues and study team members without formal involvement of funders or any other agencies. The advent of "open research" and requirements by funders that researchers develop concrete plans for data sharing makes the phenomenon to be gaining popularity. If standardized, data sharing increases statistical power and also facilitate test of new hypotheses, secure long term storage of research data (including metadata, protocols, data dictionaries) in useful formats, enhances accountability and compliance with best practices. There are public repositories for data sharing in many developed countries. Some journals also make it mandatory to deposit data analysed for a manuscript in repositories.

RESPONSIBILITY AND CAPACITY BUILDING IN DATA MANAGEMENT

Just as good clinical research is a team work, likewise data management is team effort. Depending on funding and how big a research project is, data manager may be appointed. Occasionally, the study statistician takes responsibility for data quality assurance. It must be emphasized that statistical activities span every stage of the research process. From study design, to implementation, analysis and report writing. Even if roles of data manager and statistician are separated, both of them need to work with every member of a research team to ensure the collection of good quality data. Generation of valid and reliable data is an important tenet and ethical responsibility in clinical research.

The field of data management is evolving; therefore a number of standardised guidelines are being developed to guide practitioners. Examples of these include Good Clinical Data Management Practices developed by the Society for Clinical Data Management. The Clinical Data Interchange Consortium (CDISC) has also developed two guidelines: Clinical Data Acquisition Standards Harmonization (CDASH) and Study Data Tabulation Model Implementation Guide (SDTMIG). Several institutions are also offering postgraduate programmes in clinical or research data management.

CONCLUSION

Data management is the process of ensuring that data collected during a study is accurate, complete, logical and consistent. It involves verification, validation and quality control. Good data management practice is the hallmark of scientific integrity. It requires passion for excellence, attention to details, versatile skills and maximum cooperation of the research team. The processes need to be well thought-out, implemented with precision and documented for future reference.

REFERENCES

CDASH (2011). Clinical Data Interchange Standards Consortium. Available from: http://www.cdisc.org/cdash.

Coursera (2019). Data management for clinical research [Internet] Vanderbilt University; https://www.coursera.org/course/datamanagement.

Cummings J, Masten J (1994). Customized dual data entry for computerized data analysis. Qual Assur;3:300–303.

Daniel W W (2006). Biostatistics: A Foundation for Analysis in the Health Sciences, 7th Edition. Wiley, New Delhi, India.

Fegan GW, Lang TA (2008). Could an open-source clinical trial data-management system be what we have all been looking for? PLoS Med 5:e6.

Gerritsen MG, Sartorius OE, vd Veen FM, Meester GT (1993). Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation. Proc Annu Symp Comput Appl Med Care 659–62.

Haux R, Knaup P, Leiner F (2007). On educating about medical data management - the other side of the electronic health record. Methods Inf Med.46:74–79.

Kirkwood R.K. and Sterne J.A.C (2004). Essential Medical Statistics, 2nd Edition. Blackwell Publishing, Oxford, UK

Kuchinke W, Ohmann C, Yang Q, Salas N, Lauritsen J, Gueyffier F, et al (2010). Heterogeneity prevails: The state of clinical trial data management in Europe - results of a survey of ECRIN centres. Trials. 11:79.

Lu Z, Su J (2010). Clinical data management: Current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials. 2:93–105.

Ottevanger PB, Therasse P, van de Velde C, Bernier J, van Krieken H, Grol R, et al (2003). Quality assurance in clinical trials. Crit Rev Oncol Hematol. 47:213–235.

Peat J. and Barton B. (2005). Medical Statistics: A guide to data analysis and critical appraisal. Blackwell Publishing, Oxford, UK.

Read KB, Surkis A, Larson C, McCrillis A, Graff A, Nicholson J, Xu J (2015). Starting the data conversation: informing data services at an academic health sciences library. J Med Lib Assoc. 103(3):131–135.

Study Data Tabulation Model (2011): Clinical Data Interchange Standards Consortium. Available from: http://www.cdisc.org/sdtm

University of Edinburgh (2018). Research data management programme: http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-home.