Global Journal of Health Science; Vol. 13, No. 3; 2021
ISSN 1916-9736 E-ISSN 1916-9744
Published by Canadian Center of Science and Education

Data Cleaning Needs and Issues: A Case Study of the National Reproductive Health Assessment (RHA) Data from Solomon Islands

Richard D. Nair[1], Latileta L. Odrovakavula[1], Masoud Mohammadnezhad[1], K. Venkata Raman Reddy[2], Dilan A. Gohil[3] & Shiwanjani S. Sami[4]

[1] School of Public Health and Primary Care, Fiji National University, Suva, Fiji

[2] Rakiraki Hospital, Rakiraki, Fiji

[3] Colonial War Memorial Hospital, Suva, Fiji

[4] Nakasi High School, Nausori, Fiji

Correspondence: Masoud Mohammadnezhad, Associate Professor in Public Health (Health Promotion), School of Public Health and Primary Care, Fiji National University, Suva, Fiji.

Received: December 12, 2020 Accepted: January 10, 2021 Online Published: January 21, 2021

doi:10.5539/gjhs.v13n3p23 URL: https://doi.org/10.5539/gjhs.v13n3p23

Abstract

Data cleaning is an essential part of any research work without which the validity and reliability of the data could come under the spotlight. **Aim:** to document common errors found during the cleaning of datasets and suggests ways of minimizing errors during data entry process, reducing human errors throughout data cleaning.

Design and Setting: a case study based on the national Reproductive Health Assessment (RHA) data conducted in Solomon Islands in 2013.

Objective: The main objective of the Solomon Islands RHA was to establish the health status of reproductive aged women between the ages of 15 – 49 for the Solomon Islands.

Method: Data was collected using questionnaires and entered on to the SPSS database in the country by the local Solomon Islands research assistants who were trained by the Pacific Sexual and Reproductive Health Research Center (PSRHRC). The data was brought back to Fiji where the cleaning process took place.

Results: Findings of this case study showed that there were issues with the standardization of databases, database familiarization and data merging.

Conclusion: More training is needed for researchers who are involved in data collection, data entry and data cleaning to minimize such errors which could give results which may not be a true representation of the indented study.

Keywords: data cleaning, data entering, error, research education, standardization

1. Introduction

Data collection and data analysis have been carried out with obligatory importance for most research work. However, quality of data remains a concern as the presence of incorrect or inconsistent data can significantly distort the outcomes of analyses and incorrectly inform results (Chapman, 2005). The quality of data generated plays an important role in research (Krishnankutty, Bellary, Kumar, & Moodahadu, 2012). According to Brown, Kaiser, and Allison (2018), poor quality data refers to those acquired through erroneous or sufficiently low - quality collection methods, study designs, or sampling techniques, such that their use to address a particular scientific question is scientifically unjustifiable. Cai and Zhu (2015) further elaborated that serious decision-making mistakes would be endured if the data quality is poor which will further lead to low data utilization efficiency

Data cleaning follows the processes of data collection (field work) and data entry after which data is thoroughly checked for errors and other inconsistencies are corrected before analysis begins (Rahm & Do, 2000). Even though the importance of data-handling procedures is being underlined in good clinical practice and data management guidelines, gaps in knowledge about optimal data handling methodologies and standards of data quality are still present (Van den Broeck, Cunningham, Eeckels, & Herbst, 2005). According to Rahman, Desa, Wibowo and Haris

(2019) a huge body of research has dealt with schema translation and schema integration; however, data cleaning process has received only little attention in the research community. Errors of data management tend to be more idiosyncratic than systematic, with most errors being due the construction of bespoke methods of handling, storing, or otherwise managing data (Brown, Kaiser, & Allison, 2018). Rahm and Do (2000), further elaborate that data quality problems are present in single data collections, such as files and databases due to misspelling during data entry, missing information, or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly.

To address data quality issue, there is ample literature that has discussed various aspects of data cleaning, highlighting the importance of the process for validating the existence of good quality data. According to Cai and Zhu (2015) appropriate quality assessment method for big data is necessary to draw valid conclusions the authors, further state, that in order to further quality assessment, specific assessment indicators for every dimension must be chosen, and thus the data needs to such that it complies with specific conditions or features. This includes frameworks and guidelines that depict steps of cleaning data. Frameworks highlighted include defining and determining the error types, searching, and identifying error instances, correcting the errors, documenting error instances and error types; and modifying data entry procedures to reduce future errors (Maletic & Marcus, 2000). Rahm and Do (2000) further elaborate that to support data quality detailed information about the transformation process is to be recorded, both in the repository and in the transformed instances, in particular information about the completeness and freshness of source data and lineage information about the origin of transformed objects and the changes applied to them. Data cleaning can be divided into four patterns based on implementation methods and scopes, these are: manual implementation, writing of special application programs, data cleaning unrelated to specific application fields, and solving the problem of a type of specific application domain (Harvey, Zhang, Nixon, & Brown, 2007). Guidelines and procedures are available in guiding health researchers in conducting efficient data cleaning but, contrary to expectations, there are still many loopholes that cause inconsistencies and inaccuracies in the entries for analysis (Rahm & Do, 2000).

Statistical software's adequately provide computational techniques in trying to and in some cases correct errors in data but for most cases data can only be effectively cleaned by human efforts. Humans are usually involved in explanatory data mining in understanding the dataset, identifying errors and the rectification of these errors (Hellerstein, 2008).

Manual process of data cleaning involves laborious efforts and is time consuming and is itself prone to errors (Maletic & Marcus, 2000). This paper aims to document errors found during the cleaning of datasets and suggests ways of minimizing errors during data entry process, reducing human errors throughout data cleaning. For this paper, data cleaning examples are drawn from cleaning the 2013 Solomon Islands Reproductive Health Assessment (SIRHA) data. Solomon Islands is the third largest archipelago spread over 900 islands that is divided into nine provinces, the country has a population of approximately 0.6 million with diverse language and culture (Solomon Islands Office | UNDP in the Pacific, 2020)

1.1 A Brief Introduction of the SIRHA Study

The 2013 Solomon Islands Reproductive Health Assessment (SIRHA), was a national undertaking of the United Nations Population Fund (UNFPA), Pacific Sub-Regional Office (PSRO). The UNFPA commissioned the Pacific Sexual and Reproductive Health Research Centre (PSRHRC) to implement the assessment in partnership with, local stakeholders, including the Solomon Islands Ministry of Health & Medical Services (MHMSs), Solomon Islands College of Higher Education (SICHE), National Statistics Office (NSO) and the Solomon Islands National Planned Parenthood Association (SIPPA) (Rokoduru, 2014)

The main goal of the SIRHA was to establish the health status of reproductive aged women between the ages of 15–49 for the Solomon Islands. To achieve this, a sample of randomly selected households with an average number of 3 reproductive-aged women per household was selected for the assessment from the three largest and most populated provinces namely Malaita Province, Guadalcanal Province and Western Province (Rokoduru, 2014)

The assistance of NSO was sort in partnership with the Fiji Islands Bureau of Statistics (FBoS), the representative sample population of 800 households was calculated in line with the available budget from UNFPA PSRO. Using standard selection procedures and formulas for non-biased selection of enumeration areas and sample population for the SIRHA, the project was implemented in a total of 40 enumeration areas in the three provinces with each of the enumeration areas yielding randomly selected 20 households for this survey (Rokoduru, 2014)

2. The Need for Data Cleaning

The need for data cleaning is based on improving the quality through reducing errors in data and improving their documentation and presentation. Chapman (2005) in his report stated that, unless extraordinary efforts are taken, a field error of 1-5% is to be expected. From data cleaning experiences of research data, it has been identified that these inaccuracies are mostly attributed to human error during database designing, data entry and even data cleaning. Lack of knowledge and purely unpremeditated error are some of the reasons of incorrect entry of data and cleaning of data (Chapman, 2005). According to Rahman et al. (2019) data cleaning problems can arise in single-source and multi-source problems and between schema- and instance-related problems. Schema – level problems can be addressed at the schema level by an improved schema design, instance – level problems, on the other hand, refer to errors and inconsistences in the actual data content which are not visible at the schema level.

Rahm and Do (2000) further elaborated that single source problem occurs when the data quality of a source largely depends on the degree to which it is governed by schema and integrity, for instance, files, have few restrictions on what data can be entered and stores, giving rise to a high probability of errors and inconsistencies. Multisource problem, on the hand, exists when multiple sources need to be integrated for data, and each source may contain dirty data and the data in the sources may be represented differently, overlap, or contradict.

The data entry phase is the most prone for occurrences of errors as compared to any other phase of data management. Rahman et al. (2019) states that, data entry phase encompasses many errors due to misspellings, missing information, or even invalid data. This is because the sources often contain redundant data in different representation. Higher incidences of errors are to be expected if there are multiple data entry personnel (Rahm & Do, 2000). An example is the data entry for the SIRHA whereby data entry was done in 9 laptops and the three provinces studied were entered into separate databases in these laptops. Laptops were assigned to individual data entry personnel's and there were cases where more than one individual entered data in a laptop. Evidently enough, there errors were observed in the data which required ample time and effort for data cleaning.

For the data cleaning phase, errors may occur, and this is more evident when working with large databases. However, most of these errors are based on lack of understanding and imprecision of the data cleaning personnel. Similarly to data entry, having multiple people clean a dataset with the absence of mutual understanding of the database and its requirements of data cleaning, can create more inaccuracies (Rahman et al, 2019). The errors that may occur during data entry and data cleaning are further discussed in the following topics.

3. Standardization of Data

To ensure research data is recorded and presented correctly, most research rely on accurate developed databases, these are either consolidated or standalone (Cross, Palmer, & Stephenson, 2009). To allow this, the concept of standardization is to be applied, this concept aims to reduce errors in research (Nissinboim & Naveh, 2018). Standardization in research, often refers to, ensuring that processes in research are kept the same, this considers the research methods to the definition and the measurements of variables of interest (Salkind, 2010). The concept requires all elements of the database to be planned and developed accordingly. Standardization of database designs and data entry is further discussed in this section of the paper.

3.1 Standardization of Database Design

To aid researchers' management of research data, numerous data software is increasingly being developed (Zamawe, 2015). These software's enhance data management process allowing data to be managed from its rawest form to analysis (Surkis & Read, 2015). To ensure that the correct analysis results, human data entry needs to be scrutinized and data entry conducted properly. A single incorrect data entry can turn a significant test to non-significant (Barchard & Pace, 2011). The standardization concept allows human error to be reduced, as it ensures that data to be entered in certain formats. Data cleaning applied to the SIRHA data reiterated the importance of standardization of the database design as data was entered into more than one database. The Statistical Package for the Social Sciences (SPSS) was the statistical software used for SIRHA study. SPSS software allows users to design databases according to specific set of questions and provide analytical processes for generating results (Hinton, 2014). There are two views on a SPSS database; (i) variable view, for designing of variables and (ii) data view, where data is entered and stored. Data was entered into SPSS version 20 by data assistants at Solomon Islands. Excel spreadsheets were also used for the cleaning of the datasets. The SPSS variable view allows measures and specification to be applied to variables. Examples of application of standardization of databases from the SIRHA database are, database designs to accept alphabetical answers only for string variables and numbers for numeric variables, allocation of values for coded variables, limitation to the length and width of characters entered, number of decimal places and assigning measures to data whether it be nominal or scale.

For the SIRHA database, data cleaning highlighted the need to format the database design to align the requirements of data entry accordingly for all the databases considering the above factors of assigning certain measures to each variable. For example, the variables can be either marked string which allows alphabetical answers to be entered or numeric which only allows data to be entered as numbers. The age variable in most instances were marked as either string, or numeric which requires the data entry personnel to either fill in the age as “54” for numeric or “54 years” for string causing anomalies during the merging of the datasets. Age variables were also shown as 54.0 and 54.00 for databases that have assigned decimal places to the age variables. The coding of variables was inconsistent; some databases had assigned a string measure or allowed alphabetical answers to be entered allowing databases to have either ward codes as 1 to 58 or ward names. Appropriate labeling of variables and the assigning of correct measures is to be established when designing a database. In scenarios where more than one individual is involved in cleaning data, it is paramount that a standardized data set is used by all, with mutual understanding of standardized formatting and entry.

Apart from applying measures to databases, it is vitally important to ensure developed databases capture intended information and it is important that this be supported by the database design. For the SIRHA, variables were seen to be missing from some of the databases and this was addressed through the insertion of both the variables and its values. This has been attributed to human error, either in database designing or data entry. It is recommended that databases be thoroughly checked, before it is given to data entry personnel for data entry. The best ways in preventing many errors is to professionally design the database

3.2 Standardization in Data Entry

Observations made from data cleaning experience is that although the design of a database controls the level of standardization there is a fair input data entry on the existence of standardization. This observation was drawn from the SIRHA dataset. For this assessment data entry was conducted in-country with the main idea to build capacity and develop research in areas of data collection and data entry. With the magnitude of the assessment, it was ensured that data entry personnel’s received training on the reproductive health assessment and the requirements of entering the collected data. However, errors and the lack of standardization were observed in the dataset and this is to be expected as literature has highlighted that typically around 5 % errors is to be expected in large databases unless extreme measures are applied to safeguard data entry processes (Maletic & Marcus, 2000). Data entered lacked consistency in the formats in which they were entered, and this was observed for all the datasets. Drawn observed examples are questionnaire code numbers which were the identification numbers were of different lengths with total of 12, 15 or 16 digits, months and village names and were either entered with full names, shortened or misspelt. In reference to the SIRHA study, it can be inferred that large data sets, will require maximum effort in ensuring that standardization begins from initial data entry phases. Efforts in the forms of proper training of individuals involved in data entry.

4. Database Familiarization

From the data cleaning experiences with the SIRHA, it had been identified that possible factors that contribute to data entry and data cleaning is the unfamiliarity and the lack of knowledge of the study database. It can be inferred that, not only do data entry personnel’s need to familiarize themselves with the study variables, but it is also vital for them to understand the questionnaire itself considering elements such as codes and values. Lack of proper data entry and thorough data cleaning will arise if data personnel’s do not fully understand a data set including skips and filters (Van den Broeck et al. 2005). This has been noted during data cleaning, whereby skip questions have been answered and entered in the dataset, creating confusion amongst personnel’s responsible for data cleaning who tried to figure out what the correct values were for the responses depicted in the dataset.

Training sessions should discuss clearly on the database and the variables, definitions, values, and labels. Adequate emphasis is to be made on the formatting of data during entry for the purpose of standardization. Training sessions should also familiarize data entry personnel on the database increasing their knowledge of the data allowing them to safe check values they are entering. An anomaly noted in the dataset was that data assistants responsible for entering data did not know the difference between the coding 97 and 99 and used this coding’s at the wrong places.

5. Merging Data Sets

In dealing with large databases, usually more than one data entry personnel are involved and work on individual databases warranting the need for merging of large datasets and this has proved to be problematic for most cases (Liu, Simon, Amann, & Gançarski, 2020). This was the observation made during the merging of the databases for the SIRHA. Lack of consistency and standardization of the databases did not allow quick and efficient merging of the databases, this was also highlighted by a study done by Voss, Ma, and Ryan (2015) which explored the overall prevalence of inpatient conditions in the raw data and comparing to that in the Common Data Model (CDM) with applying the standardized visit definition, they concluded that having standardization of databases reduces the heterogeneity. The lack of consistency of databases for the SIRHA was addressed by amending individual databases which required both time and effort.

Merging of large datasets also has its risk due to the magnitude of the data. In relation to the SIRHA data, merging had to be well organized and structured due to the numerous databases that required merging. As noted in the merging phase of this database, there is possibility of data being mixed-up due to functions of software used. Empirical data suffer from many types of errors, one of which is called a standard merging error, recognizing various individuals as one entity. Relevant author's features were used to eliminate mistakes in merging. Xie (2019) suggested a Bayesian model on the fusion of co-authorship data errors. The model contributes to finding informative features to reduce the merging errors of the data sets obtained by the same method when knowing the ground truth of specific empirical datasets obtained using a given method. The model can be used to measure the rate of merging errors for the naming entities of authors when given the useful features of reducing merging errors. The model can therefore help identify compromised naming entities; hence it has potential contribution to enhancing the quality of empirical co-authorship data.

Proper understanding of software used is required for data cleaning personnel. This can be achieved through training of data personnel on data management which is inclusive of the importance of correct data entry, cleaning, and dataset merging. A study by Abubakar, Atala, Abdullahi and Abdullahi (2019) also concluded that possession of training, education and job experience including having numeracy skills (understanding mathematics and statistics) are the most important requirement for an effective data management. Training, education, and experience were found to be positively and statistically related to data management requirements of extension personnel in Kaduna.

Another factor to consider when merging data sets, is the development of a standardized template to be used by all data personnel. Changes made to variables at any time during data processing, needs to be documented so that all the other datasets could be amended accordingly for efficient merging process.

6. Conclusion

Errors are bound to occur for research data, and these are removed by the processes of data cleaning as the practice aims to perfect any data set before the analysis process. Standardization, familiarization, and data merging are three fundamental principles to be considered in databases. With the existence of these principles, data cleaning, merging and analysis can be done efficiently and quickly.

Errors in databases distort data, misrepresenting findings, leading to questionable analysis and results. High occurrences of errors can be time consuming and strenuous efforts are often required for large cleaning of large databases. This may affect research completion deadlines, as the rectification of merging problems takes a considerable amount of time to iron out. Obstacles that shadow the process of data cleaning need to be looked at from a bird's eye view, with issues such as standardization of the template used for entry and errors in data entry. Errors in data entry, cleaning and merging, could distort many vibrant and promising research works to the wire, raising many questions as to the validity of the research work.

It is highly recommended that training of personnel who are involved from the initial stage of data entry, data cleaning and even data merging and analysis receive adequate training to allow proper understanding of the processes therefore decreasing the chances of high error occurrences. Possible methods to further reduce errors for research data need to be explored and applied to strengthen data quality.

Competing Interests Statement

The authors declare that there are no competing or potential conflicts of interest.

References

Abubakar, L., Atala, T. K., Abdullahi, H. A., & Abdullahi, J. A. (2019). Data Management Capabilities of Extension Personnel in Kaduna Agricultural Development Agency Kaduna State, Nigeria. *Journal of Agricultural Extension, 23*(4), 22-29. https://doi.org/10.4314/jae.v23i4.3

Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior, 27*(5), 1834-1839. https://doi.org/10.1016/j.chb.2011.04.004

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences, 115*(11), 2563-2570. https://doi.org/10.1073/pnas.1708279115

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal, 14*(0), 2. https://doi.org/10.5334/dsj-2015-002

Chapman, A. D. (2005). Principles and Methods of Data Cleaning - Primary Species and Species Occurrence Data, version 1.0. *Report for the Global Biodiversity Information Facility, Copenhagen*.

Cross, S. S., Palmer, I. R., & Stephenson, T. J. (2009). How to design and use a research database. *Diagnostic Histopathology, 15*(10), 490-495. https://doi.org/10.1016/j.mpdhp.2009.07.003

Hinton, P. R. (2014). *Statistics explained*. https://doi.org/10.4324/9781315797564

Harvey, A., Zhang, H., Nixon, J., & Brown, C. J. (2007). Comparison of data extraction from standardized versus traditional narrative operative reports for database-related research and quality control. *Surgery, 141*(6), 708-714. https://doi.org/10.1016/j.surg.2007.01.022

Hellerstein, J. M. (2008). Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE), 25*. http://doi.org/ 10.1.1.115.6419

Krishnankutty, B., Bellary, S., Kumar, N. B., & Moodahadu, L. S. (2012). Data management in clinical research: an overview. *Indian journal of pharmacology, 44*(2), 168. https://doi.org/10.4103/0253-7613.93842

Liu, R., Simon, E., Amann, B., & Gançarski, S. (2020). Discovering and merging related analytic datasets. *Information Systems, 91*, 101495. https://doi.org/10.1016/j.is.2020.101495

Maletic, J. I., & Marcus, A. (2000). *October. Data Cleansing: Beyond Integrity Analysis*. In Iq p.200-209. https://doi.org/ 10.1.1.37.5212

Nissinboim, N., & Naveh, E. (2018). Process standardization and error reduction: A revisit from a choice approach. *Safety science, 103*, 43-50. https://doi.org/10.1016/j.ssci.2017.11.015

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*(4),3-13. https://doi.org/ 10.1.1.98.8661

Rahman, F., Desa, M., Wibowo, A., & Haris, N. (2019). Data Cleaning in Knowledge Discovery Database-Data Mining (KDD-DM). *International Journal of Engineering and Advanced Technology, 8*(6S3), 2196-2199. https://doi.org/10.35940/ijeat.F1100.0986S319

Rokoduru, A. (2014). *Annual Progress Report 2013*. Pacific Sexual & Reproductive Health Research Centre.

Salkind, N. J. (ed., 2010). *Encyclopedia of research design* (Vol. 1). Sage. https://doi.org/10.4135/9781412961288

Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association: JMLA, 103*(3), 154. https://doi.org/10.3163/1536-5050.103.3.011

UNDP. (2020). Solomon Islands Office | UNDP In The Pacific [online]. Retrieved from https://www.pacific.undp.org/content/pacific/en/home/about-us/soi-office.html

Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med, 2*(10), p.e267. https://doi.org/10.1371/journal.pmed.0020267

Voss, E.A., Ma, Q., & Ryan, P. B. (2015). The impact of standardizing the definition of visits on the consistency of multi-database observational health research. *BMC medical research methodology, 15*(1), 13. https://doi.org/10.1186/s12874-015-0001-6

Xie, Z. (2019). A Bayesian model on the merging errors of coauthorship data. *Physica A: Statistical Mechanics and its Applications, 527*, 121140. https://doi.org/10.1016/j.physa.2019.121140

Zamawe, F. C. (2015). The implication of using NVivosoftware in qualitative data analysis: Evidence-based reflections. *Malawi Medical Journal, 27*(1), 13-15. https://doi.org/10.4314/mmj.v27i1.4

Copyrights