BIOSTATISTIC SERIES

Statistics Corner: Data Cleaning-I

Kamal Kishore[1], Rakesh Kapoor[2], Amarjeet Singh[3]

Reality check

"Let us assume that an investigator collected various demographic, clinical, psychiatric, and radiological characteristics of the study participants." The investigator took adequate precautions to enter data in a structured format into a spreadsheet. However, before proceeding ahead, the investigator wanted to ensure that data are ready for analysis. In this context, the investigator reviewed the literature and came across the term "data cleaning." The fellow colleagues advised him to approach a statistician for cleaning and analyses of the data. The investigator was in dilemma, whether to share the data with a statistician before or after cleaning. The investigator reviewed the literature and found some answers regarding the role and responsibilities of the investigator in data cleaning. However, the investigator still had the following questions for data cleaning.

- Is data cleaning practice a part of good clinical practice (GCP)?
- Is it the responsibility of a statistician to clean and code the data?
- Do data cleaning begin after data entry?
- How to deal with missing values at the data entry stage?

Journal of Postgraduate Medicine Education and Research (2019): 10.5005/jp-journals-10028-1330

Introduction

A majority of investigators from medical sciences collect study data in a hard format from the patients. These data are subsequently entered manually in a spreadsheet or statistical software for appropriate analysis and interpretation. Despite taking utmost care during study design, data collection, and entry stages, errors and inconsistencies are unavoidable. These errors range from variable names, variable codes, impossible data values, missing data values, etc. to name a few. These errors and inconsistencies can be corrected for making meaningful use of the data through data cleaning, a crucial initial step before analysis. An unclean data may lead to a combination of a host of problems such as wastage of time, missingness of subjects and variables, inconsistencies, outliers, and wrong results. As per estimates, 80% of the time from data entry until analysis goes in data cleaning.[1] Despite data cleaning being repetitive and major time consumer, there is little research in efficiently cleaning the data.[2] Moreover, data cleaning requires collective efforts from the study investigator and statistician to clean and code the data. Data cleaning is one of the most neglected areas of research. However, due to the increasing adoption of GCP guidelines and regulations, the importance of data cleaning practices is on the rise.[3] Moreover, of late, there is an increasing thrust by journals and sponsoring agencies to submit data in public data repositories or along with reports and articles. Now, digital object identifier (DOI) is also separately generated for the data which may fetch citations to the author. Thus, there are compelling reasons to efficiently clean and code the data for submission. Replication and validity are the hallmark of the well-documented and conducted studies. The growing clamor among various leading journal articles[4–7] highlights the need for rigorous conduct, capture, clean, and availability of dataset in public repository. In this regard, the American Statistical Association (ASA) has recommended to make it a standard practice of reporting data cleaning description along with a statistical plan of analysis.[8] The investigators should be careful about errors and inconsistencies in the data so that the data cleaning process is smooth. These can enter into the studies from multiple phases and sources. It can be segregated into human and machine errors, before and after the data entry stage, and so on so forth.

[1,2]Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

[3]Department of Community Medicine, 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

How to cite this article: Kishore K, Kapoor R, *et al*. Statistics Corner: Data Cleaning-I. J Postgrad Med Edu Res 2019;53(3):130–132.

Source of support: Nil

Conflict of interest: None

We will be broadly segregating these errors and inconsistencies in four phases of research which are presented in Table 1. These four phases can be further segmented into preventive-cleaning (phases I and II) and active-cleaning (phases III and IV) phases. These phases are discussed subsequently. However, it is not an exclusive and exhaustive list. It is a general set of principles which should be followed routinely. Investigators of studies should carefully think about research implications and resources at hand to take precautionary measures to minimize errors and inconsistencies.

Phase I

Data cleaning begins with the study design phase. This phase of the study is very crucial for completeness and data capture of the variables. Investigators should carefully plan about the variables and measurement scales to collect them. It should be noted that a quantitative variable can be transformed and categorized at any stage of study duration but the same is not true for a categorical variable. Investigators should carefully think to measure variables like education in a quantitative (number of years of formal education) or categorical (primary, high school, college, etc.) scale. Similarly, the timing of follow-ups in numbers, hours, days, months, year, ages, etc. should be carefully decided in the beginning. We have come across poorly defined examples of variables where it becomes difficult to code at

Table 1: List of activities to attain clean data for statistical analysis

Phase	*Error specification*	*Actions required at each phase to clean data*
Phase I (study design)	Variable measurement	Collect at highest measurement scale, Quantitative ⇄ Categorical
	Follow-ups	Measure and enter as per dates of data collection
	Variable classification	Decide and prepare a list of combinable labels in a variable
	Multisite study	Generate unique IDs, standardize data entry
	Missing data	Plan and document the possible reasons for missing data
Phase II (data collection)	Continuous variable	Use reliable and valid instrument of one manufacturer
	Duplication	Maintain and update separate list for units of data collection
	Missing data	Carefully capture the reasons for no answer
	Check boxes	Restrict to allocated space for the check boxes and tick marks
	Ill-defined question	Avoid ambiguous, loaded, double barrel*, etc. questions
Phase III (data entry)	Variable label	Restrict to 8–10 characters, use underscore as separator
	Value label	Maintain uniformity of codes
	Extreme value/outliers	Carefully assess and cross-check the outliers
	Unique ID	Generate and use artificial generated unique IDs
	Missing value	Use artificial coding such as −111, −999, etc. for missing values
Phase IV (data processing)	Data export	Export appropriate sheet, inspect variables and cases for errors
	Mathematical operations	Retain the codes for conversions or/and recoding
	Classification	Classify variables into appropriate measurement scale
	Coding	Code appropriately to avoid duplication and identification
	Select cases	Carefully select and unselect cases as per the objectives

*A double-barrelled question address two or more that two separate issues or topics (Is this article interesting and useful?) but it can only have one answer

later stages during cleaning and analysis. An example of this could be occupation where professional, semi-professionals, business, and other categories are not properly defined (e.g., Whether a Dr who owns a clinic will be labeled under professional or business category?). The message is to clearly think of how and what type of categorization is apt for the study outcomes. Multisite studies are on the rise, it brings more complications on the table. There is a lot of variations in data entry and cleaning practice as per academic qualification and training. Therefore, investigators of multisite studies should prepare a standard entry, cleaning, and coding document. It should be used as a reference to enter, clean, and process the data for analysis. Moreover, a unique identification (ID) and process to compile all data should be carefully deliberated in the beginning by an investigator and a statistician. This initial process and document will save a lot of time and hassle at the later stage of data processing. Despite best efforts, missing values in the data are not exceptions. It is important to segregate among "don't know," "refuse to answer," "not collected," "don't care," etc. categories and it should be carefully discussed at the beginning of the study. A look at the entered data, in the beginning, can highlight deficiencies which can be corrected at initial stages of data collection.

Phase II

Data collection is important and is the soul of study. Therefore, adequate planning, training, and care should be taken to capture the data. The difference in physical measurements such as waiting time, height, or weight recorded with different instruments in haste may lead to a difference in observations due to instruments and investigators biases. Therefore, the use of standard valid instruments of one making is advisable. Similarly, inadequate time to each participant for filling the information may lead to a drop in quality and quantity of data collection. Once information is collected from the participants, the same should be carefully stored in a safe place. It will be useful at later stages for data entry and validation. Duplication and loss of forms also happen; therefore, each form should be given a unique ID. A couple of times, it is difficult to read the handwriting. Therefore, data collectors should be trained and emphasized to collect data in lucid and easy to read the language. It may save time and errors while entering the data. Similarly, tick marks ($\surd$), cross marks ($\times$), circles, etc. should be limited to allocated space in the form. The extension of tick marks, etc. beyond allocated space can create confusion which may lead to delay and errors during data entry in a spreadsheet or software. Many a times, vague questioning (Is this article interesting and useful?) suffer from no response or multiple responses. Investigators should aspire to capture only one attribute in one question (Is this article interesting?). It should be discussed among the investigators to correctly frame the questions in an objective way.

Phase III

Data entry is the process in which collected data are entered in the appropriate spreadsheet or statistical software to extract information. The errors in this stage can be broadly segmented into cosmetic and logical errors. Cosmetic errors involve renaming of variables, labels, detecting outliers, etc. Typically, a variable label should be 8–10 character long (e.g., What is your date of birth can be coded as DOB) without any spacing in between. Further, it is recommended to use underscore (_) for separate words as compared to other special characters (#, $, &, etc.). More information regarding variable naming and data structure can be obtained from "Structured Data Entry" in the "Biostatistics Series."[9] The value labels for categorical data should be uniform as most of the software are case sensitive (e.g., either all males should be coded as Male or MALE but not some as Male and some as MALE). Similarly, a continuous variable can be validated for extreme values by arranging data in increasing and decreasing orders. It will give a quick peek into extreme lower or higher value which may warrant further inspection and action. It is very important

to understand the importance of random unique ID variable at this stage. Avoid clinic number, registration number, or any other unique ID generated under health setup so as to prevent identification of patients. Moreover, the data go through repeated filtering, cleaning, and ordering stages. Thus, the unique ID can be used to revert data into the original sequence. However, the major advantage of the unique ID is to trace the patient record to the original file which may be used to confirm, code, or change the value in case of discrepancy. Missing values can be coded as −999 or −111 or any other numbers with a negative sign as most of the observations from patient have positive values.

However, it is not a universal rule and care should be taken on a study-to-study basis. Data deletion and duplication may occur unknowingly at the entry stage. Run frequency analysis on the unique ID to know the status of duplication and deletion. A logical error occurs in a situation where value labels of one variable are inconsistent with another variable. It is highly unlikely for a child of 10 years to have a Ph.D. degree. Similarly, males cannot be pregnant. These kinds of inconsistencies in data come under logical errors. These are usually assessed with simultaneous application of filter option or cross-tabulation of two variables.

Phase IV

Finally, errors in this phase can be attributed to miscellaneous or machine category or data-processing errors. As per our experience, the majority of the investigators first enter data in a spreadsheet before exporting it to appropriate software for further analysis. Sometimes, after exporting, the data display extra columns with variable names V1, V2, etc. at the end of the variable column which are not a part of original study variables. Similarly, extra rows with dots (.) for missing values can appear in cells. A quick run of frequency analysis of unique ID and variables can indicate these anomalies in the form of either high or low number of participants and variables than originally planned in the study. Delete unwanted empty rows and columns. A recoded variable obtained by merging of labels (e.g., two religious categories Buddhism and Jainism are merged to form a new label as other religious groups) and transformations (log, square root, BMI calculations, etc.) are stored as a new variable without deleting original variables. It is a good practice to add R or Re as prefix or suffix to identify recoded variables. Similarly, some survey questionnaire requires reverse coding for analysis and meaningful interpretation. Create and code new variables by adding reverse/R/Re as suffix or prefix to original variables. Since most of the statistical software's require observations from both quantitative and categorical variables to be stored in numbers. Therefore, categorical variables are classified after exporting the data to appropriate software for analysis. Broadly, exported variables in the data are declared as strings for names, ID, etc., nominal for gender, religion, ordinal for ranking, and scale for quantitative variables. Similarly, there are other classifications available for the variables. The knowledge of measurement scales is a prerequisite to code and clean the data. For more detail, the interested reader can read "Measurement Scales" published under "Biostatistics Series."[10] Therefore, investigators should carefully think about the formats of the variables. Many values are edited in the data during the processing of data. The editing of values is variable (a value 1 can be used to code both male and occupation) specific.

Table 2: Typical data cleaning assessment mechanism

Characteristic	*Methods to detect errors and inconsistencies*
Data validity	Double data entry
Initial data check	Visual inspection to detect errors and inconsistencies
Unique ID	Frequency analysis for data completion, duplication and missing cases
Value labels	Frequency analysis for uniformity and selection of adequate number of labels
Quantitative	Sorting in increasing and decreasing order to detect extreme values
Inconsistency	Cross tab and filters to assess cosmetic errors and joint frequency distribution
Distribution	p–p, q–q plots and histograms to visualize normality
Outliers	Boxplot to visualize distribution of extreme values and outliers

Therefore, an investigator should carefully select and clean a single variable instead of multiple variables simultaneously. The typical mechanism to clean the data is given in Table 2.

Conclusion

Data cleaning is an important and integral part of data analysis. The process of data cleaning can be broadly segmented into preventive and active cleaning. The role of an investigator is crucial in preventive cleaning, whereas the role of a statistician is important in active cleaning. Therefore, a coordinated effort between an investigator and a statistician is crucial to achieve desired objective.

Acknowledgment

The authors would like to thank Dr Dipankar, Dr Rahul, and Dr Naveen for reviewing this article.

References

1. Dasu T, Johnson T. Exploratory Data Mining and Data Cleaning. New Jersey: John Wiley and Sons, Inc; 2003.
2. Wickham H. Tidy Data. J Stat Softw 2014;59:1–23.
3. Den Broeck JV, Cunningham SA, et al. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Med 2005;2: 966–970. DOI: 10.1371/journal.pmed.0020267.
4. Errington TM, Iorns E, et al. Science forum: An open investigation of the reproducibility of cancer biology research. Elife 2014;3:e04333.
5. Johnson VE, Payne RD, et al. On the Reproducibility of Psychological Science. J Am Stat Assoc 2017;112:1–10. DOI: 10.1080/01621459.2016.1240079.
6. Collaboration OS. Estimating the reproducibility of psychological science. Science 2015;349:aac4716. DOI: 10.1126/science.aac4716.
7. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature 2012;483:531.
8. American Statistical Association. Ethical Guidelines for Statistical Practice https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx (2018).
9. Kishore K, Kapoor R. Statistics Corner: Structured Data Entry. JJ Postgr Med Edu Res 2019;53:94–97.
10. Kishore K, Kapoor R. Statistics Corner: Measurement Scales. J Postgr Med Edu Res 2019;53:46–47.