BIOSTATISTICS SERIES

Statistics Corner: Structured Data Entry

Kamal Kishore[1], Rakesh Kapoor[2]

Reality check

"Let us assume that an investigator conducted a trial and collected various demographic, clinical, psychiatric and radiological characteristics of the participants enrolled in the study." An immediate challenge in front of the investigator is to carefully arrange data before analysis and interpretation. In this context, investigator searched the literature and found some useful tips to prepare data sheet for analysis. However, the investigator has certain questions regarding data entry for which she needs clarity.

- Is there a structured way or format to enter the data in spreadsheet or data analysis software?
- Does the structure or format of data entry vary according to the study design?
- Do nature (scale) and classification (independent or dependent variable) of variable impact data entry and analysis?
- Is there any systematic way to arrange inter and intra-clinical profile of patient?

Journal of Postgraduate Medicine, Education and Research (2019): 10.5005/jp-journals-10028-1322

Introduction

Data is routinely collected by the investigators while conducting the studies. This data is subsequently analyzed and interpreted to make meaningful conclusions. After the data is collected many intermediate steps (data entry, cleaning, coding, coding sheet, assumption validation etc.) are undertaken to make it adaptable for exporting and analysis into an appropriate software. Data entry and cleaning into appropriate spreadsheets or statistical software are crucial initial steps to efficiently store and standardize data. It lays the foundation for retrieval, understanding, and smooth analysis of the data. In this article, it is assumed that the investigator collected data in offline mode and uses Microsoft Excel® for data entry. Errors are an inevitable and integral part of the data entry process. Some perceived errors like extreme observations, missing values, etc.are genuine but others can be due to wrong entry format, naming, etc. and need to be carefully addressed. Therefore, an investigator should carefully arrange/clean and code the data in an appropriate format for the purpose of analysis. However, before undergoing the cleaning process for the data, users should know the appropriate format of data entry. Broadly data entry formats can be segregated into unstructured and structured types.

Unstructured Format of Data Entry

Unstructured format can also be called as "haphazard", "messy" or "untidy" data entry. This type of data structure does not give a clear idea of the sequence of data collection related to its objectives and outcomes. Moreover, it is riddled with errors and inconsistencies. As Wickham succinctly puts it "Every messy data is messy in its own way in contrast to tidy data which are all alike".[1] So, it is important to recognize that some of the discrepancies can be legitimate and valid. However, many others will reflect a measurement or data entry related errors.

These could mean incorrect labeling, coding, spacing, blank rows and/or columns, missing patterns, duplication, multiple responses, etc. These discrepancies range from mistakes due to human error, poorly designed recording systems, or simply because of incomplete importation of data from external data sources. Figure 1 display poor data entry practice for appropriate and efficient statistical analysis. Errors in the data entry are highlighted by circles. It can be observed that errors can range from variable names to incorrect data entry.

[1]Lecturer, [2]Professor and Head

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

Corresponding Author: Kamal Kishore, Lecturer, Department of Biostatistics, Postgraduate Institute of Medical, Education and Research, Chandigarh, India, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Kapoor R. Statistics Corner: Structured Data Entry. J Postgrad Med Edu Res 2019;53(2):94-97.

Source of support: Nil

Conflict of interest: None

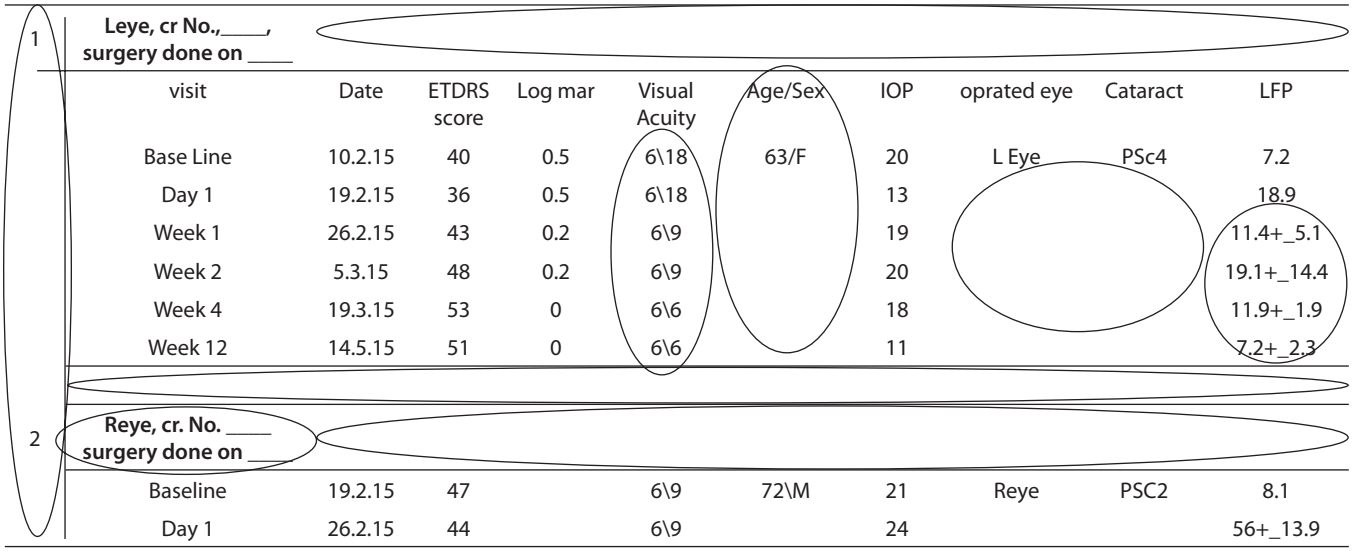
Row 1 has entered multiple variables (eye, cr no. and surgery date in a single cell of the 2nd column. It should be entered separately in each column. Subsequently, the 1st row and 1st column was left blank. Moreover, mixing of more than one variable (age/sex) and usage of symbols [visual acuity (\)] and signs (±) in the data are not the correct way of entering data. These are minor but important issues as for as data entry and analysis are concerned. Therefore, utmost care should be taken before initiating data entry into a spreadsheet or any other statistical software.

Structured Format of Data Entry

The structured format is also known as "tidy", "standardized" or "master" data entry. It is an objective and systematic way to organize the data. It organizes the data in a sequence which conforms to the objectives and outcomes related to the study. It facilitates exploration, description, and analysis of the data to attain the results. A tidy data is a rectangular array in which the row represents individuals and the column represents variables. Structured data entry clearly captures the study design used to collect the data. A cell depicts the observation for a participant corresponding to a specific variable. A variable captures the variation in a specific attribute for all the individuals. Whereas, an individual or unit presented in a row, captures all the variables. A master data sheet does not have empty rows, columns, or spacing. An illustrative master or tidy data entry format along with coding sheet is displayed in Tables 1A and 1B, respectively.

Components of Structured Data Entry

It is important to understand various components of the structured data entry. In this regard, the investigator needs to



1	Leye, cr No.,_____, surgery done on _____									
	visit	Date	ETDRS score	Log mar	Visual Acuity	Age/Sex	IOP	oprated eye	Cataract	LFP
	Base Line	10.2.15	40	0.5	6\18	63/F	20	L Eye	PSc4	7.2
	Day 1	19.2.15	36	0.5	6\18		13			18.9
	Week 1	26.2.15	43	0.2	6\9		19			11.4+_5.1
	Week 2	5.3.15	48	0.2	6\9		20			19.1+_14.4
	Week 4	19.3.15	53	0	6\6		18			11.9+_1.9
	Week 12	14.5.15	51	0	6\6		11			7.2+_2.3
2	Reye, cr. No. ____ surgery done on ____									
	Baseline	19.2.15	47		6\9	72\M	21	Reye	PSC2	8.1
	Day 1	26.2.15	44		6\9		24			56+_13.9

Fig. 1: Poor Data Entry Practices

know four fundamental requirements which are displayed in Figure 2.

A systematic array of these components is crucial to understand, retrieve and analyze. Therefore, it is important to envision the data entry framework and challenges for smooth functioning. These fundamentals are generic in nature and may not be applicable to all kinds of data. Therefore, researchers should thoroughly understand study-specific issues pertaining to their research work before commencing the data entry. An initial thought and systematic array regarding these characteristics elicit an objective, specific and organized view along with the strength and limitations of data for analysis purpose. In order to give a general idea about these characteristics, they are discussed subsequently.

Study Design

The study design is an important aspect of data collection and entry. The data entry for any study design should capture the inherent characteristics of that study. Most people are familiar with a cross-sectional data entry format. It is the simplest study design where participants are represented in a row and variables are represented by columns. It is also known as multivariate or wide format and is displayed in Table 2A.

Table 1A: Illustration of master sheet

S. No.	*Age*	*Religion*	*Family*	*Height*	*Weight*	*BMI*	*BMI_cat*	*Pain*
1.	28	2	1	155.0	90.0	37.46	5	2
2.	32	2	1	155.0	73.0	30.39	5	1
3.	26	2	1	165.0	55.0	20.20	2	1
4.	21	1	1	158.5	57.5	22.89	2	2
5.	29	2	2	155.0	64.0	26.64	4	3
6.	20	2	1	149.0	58.0	26.12	4	3

Table 1B: Illustration of coding sheet

Variable label	*Description*	*Coding and valid range*	*Measurement scale*
Serial	Serial number	None	String
Age	Age in years	None	Interval
Religion	Religion of the participant	1 = Hindu	Nominal
		2 = Sikh	
		3 = Muslim	
		4 = others	
BMI_cat	BMI classification as per	1 = Underweight (<18.5)	Ordinal
	Asian criteria	2 = Normal (18.5-22.9)	
		3 = Overweight (23-24.9)	
		4 = Obese (≥25)	

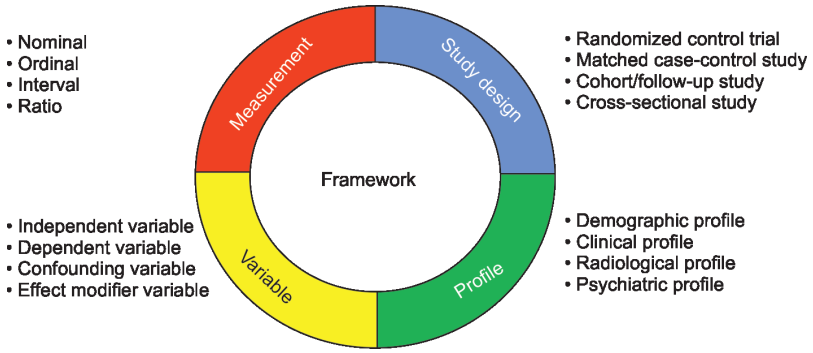


Fig. 2: Pillars of good data framework

Table 2A: Data in wide format

Subject	*Gender*	*Family*	*Edu*	*Time*$_1$	*Time*$_2$	*Time*$_3$	*Time*$_4$
1	1	0	1	36.2	39.4	42	45.3
2	0	1	0	28.6	27.5	29.4	30.6

Categorical variables entered in coded form e.g., 0→Male, 1–Female, similarly education and family are coded. Time displayed the number of follow-up of patients

Table 2B: Data in long format

Subject	*Gender*	*Family*	*Edu*	*Time*	*Neuro*
1	1	0	1	1	36.2
1	1	1	1	2	39.4
1	1	0	0	3	42
1	1	0	0	4	45.3
2	0	0	1	1	27.5
2	0	0	1	2	29.4
2	0	1	0	3	30.6
2	0	1	1	4	33.2

The data entry for a randomized control trial should identify variables belonging to the intervention and the control group. Whereas, data entry for matched case-control study design should have a matching variable to match the cases to the controls. Similarly, data entry for a cohort/follow-up study design should reflect the time points and dependent nature of data from the same participants/patients. Data for follow-up studies can be arranged both in the long/univariate format, as displayed in Table 2B and wide format. Other study designs can bring in other complexities. Therefore, researchers should carefully capture and consider the characteristic of the study design and the statistical software to be used for analysis before entering the data.

Profile

Many studies capture data from multiple clinical domains to better understand the physiological and psychological changes in the human mind and the body. The data from different domains should be systematically arranged and captured for retrieval and analysis. Intermixing of different domains may create confusion while retrieving and analysis. Moreover, the domain-wise array indicates the dependent and independent nature of variables. At this point, it should be noted carefully that some variables can act as an independent variable in one objective and dependent variable in another. Therefore, it is recommended to separate inter-profile (demographic, clinical, psychiatric, radiological, etc.) measurements from intra-profile measurements by arranging the themes in the appropriate sequence.

Measurements

There are various ways in which data can be labeled, but the most frequent labeling technique used in medical sciences is nominal, ordinal, interval, and ratio scale labeling. Ratio scale is at the top of the hierarchy of scale and offers more flexibility to the researcher for analysis. It should be noted that many of the variables such as body mass index, income, blood pressure are captured or entered categorically. It is better to capture and enter these variables as continuous variables rather than categorical as categories can change with time, place and disease conditions. Moreover, the availability of numerous statistical techniques and their usage depends on the level of measurement at which it is stored. More detail regarding the measurement scale can be obtained from the measurement scale published under Biostatistics Series.

Variables

A variable is a characteristic, which varies from person to person. The data in the majority of medical sciences are captured and stored in the form of variables. There are variables such as weight, blood pressure, education, etc. which are modifiable whereas gender, race, blood group, etc. are non-modifiable. It is always better to know the different types of variables and arrange them accordingly while entering. Variables can be broadly categorized into four categories (Table 3).

Table 3: Definitions of different type of variables

Independent variable: It is also known as predictor or influencer variable. These variables are manipulated by the experimenter to study effect.
Dependent variable: It is also known as criterion or response variable. These variables are affected by the independent variables.
Confounding variable: It is a variable which is associated with both independent and dependent variable and can distort the relationship between two.
Effect modifier variable: It is a variable that positively or negatively modifies the relationship between independent and dependent variable.

Conclusion

Structured data entry is crucial for analysis to avoid numerous mistakes. Every data is unique and thereby requires careful deliberation before starting the data entry. Most of the people start entering the data haphazardly without legitimate data validation which might create problems later. Therefore, it is recommended to consult a biostatistician during the conceptualization of the study and at the data entry stage rather than only for analysis/interpretation stage.

Acknowledgments

Authors would like to thank Professor Amarjeet Singh, Professor Hemant, Dr Dipankar, and Dr Siddhartha for reviewing this article.

References

1. Wickham H. Tidy Data. J Stat Softw 2014;59:1-23.