International Journal of Current Research and Modern Education (IJCRME)
Impact Factor: 6.925, ISSN (Online): 2455 - 5428
(www.rdmodernresearch.com) Volume 5, Issue 1, 2020

MISSING DATA, ITTYPES AND STATISTICAL ANALYSIS IN CLINICAL TRIALS

A. R. Muralidharan
Assistant Professor, Department of Statistics, College of Natural and Computational Sciences, Debre Berhan University, Ethiopia

Cite This Article: A. R. Muralidharan, "Missing Data, It Types and Statistical Analysis in Clinical Trials", International Journal of Current Research and Modern Education, Volume 5, Issue 1, Page Number 15-20, 2020.

Abstract:

A major problem in the analysis of clinical trials is missing data caused by patientsdropping out of the study before completion. This problem can result in biased treatment comparisons and also impact the overall statistical power of the study. To ensure the safety and efficacy in clinical trials, the results to be an unbiased, hence it is necessary to know the context about handling missing value in the given data.In general, most of the Clinic trials are involved with missing data. Especially missing data may seriously affect the results of inferences from randomized clinical trials, also if missing data are not handled appropriately. The potential bias due to missing data depends on the mechanism causing the data to be missing and the analytical methods. Therefore, for any analysis of trial data with missing values requires careful planning and attention.

Key Words: Clinical trials, Missing data, MCAR, MAR, MNAR

1. Introduction:

In recent decades, a very common and most frequently observed is that the data is with incomplete or missing values. The term "Missing data" is the meaning of missing type of information about the phenomena in which are interested. Commonly missing data hinder our ability to explain and understand the phenomena under any study. These missing data or incomplete data is frequently encountered in many disciplines. Statistical analysis with missing data has been an area of interest in the data analysis tasks.In this article the author's interest focus on the Missing data in the field of clinical trials. The reasons in clinical trials, there is an involvement of human subjects, more serious and expensive and also there is some complexity in using any human subjects again and again. Again the focus is on statistical analysis or approach of these missing values in clinical trials. For any quantitative data to be incomplete, in the sense that not all planned observations are actually made in that designs.The involvement of missing values is important in data analysis and data management, if the missed values are not properly handled leads to a bad results and cause on an inaccurate inference about the analysis. To avoid such misinterpretation, any analysis to be covered those missing value before analysis.

History:

In 1920's and 1930's work on the problem of missing data was largely confined to algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design. The reviews by Afifi and Elashoff (1996) and Hartely and Hocking (1971) gave more Ideas on this missing data context. In the last of twentieth century, general algorithms, such as E-M algorithm and data imputation and augmentation procedures combined with powerful computing resources, largely provided a solution to this kind of problem.

Meaning of Missing Data:

In any clinical trial, the missing data are the data which contains some no information or no values due to several activities of the patients and enumerators. These activities may cause, patient missing the follow-up, patients withdrawal and so on. Missing data gives biased results and it is a dangerous issue in statistical analysis. The ICH E9 guideline says that despite missing data, a trial may still be valid, provided the statistical methods used are sensible

Reason for Missing Data:

There are several many reasons to get a missing data set. Some of the reason are

- Data are missing in a data set when subjects in longitudinal studies often drop out before the study is completed. The reason for this drop out because the subject may die, moved out of the area, no longer no change or refuse to continue in the experiments
- Data are missing because of the participants refuse or do not know the answer in surveys
- Data are missing in experimental studies when the researcher is simply no able to collect an observation due to spoil of unit, bad climate or equipment fails

2. Missing Data Pattern:

A pattern or trend of any missing data gives an idea about the data to be handled with missing values from the given data. A pattern of missing data may point the location of the missing values in a potential

complete data matrix. For instance, let we consider a rectangular matrix data with rows representing subjects and columns representing variables. The row and column in the given data matrix can be sorted or ordered to get a special pattern of the give missing data.

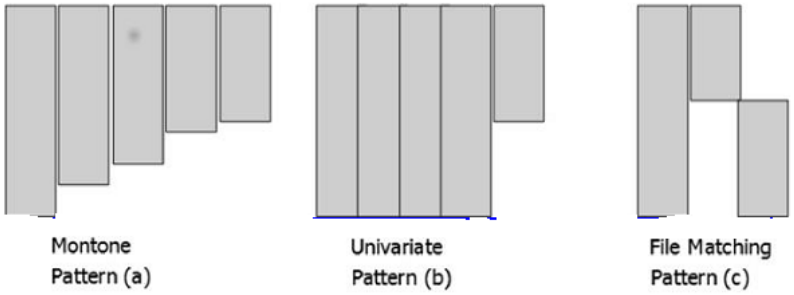

Figure 1: Patterns of Missing Data

From figure 1, the pattern of missing data were explained, there are three main patterns in missing data mechanism. Pattern (a) is the representation of Monotone pattern of missing data where the variable j=2,3,…., p is observed on a subset of subjects with variable j-1 observed points. This type of pattern of missing data arises in the cased of Longitudinal studies when the drop outs from each wave are not followed in the further times. From the figure 1 Pattern (b) is another type of pattern classified as univariate pattern, where the data is missing only on one variable in the analysis. Finally, the pattern (c) is File matching pattern, when two files are appended where F1 is file 1 and F2 is file 2, then F1 is with data on **V**1 and **V**2 and in F2 with the data on **V**1 and **V3**, then the pattern existed. This type of pattern also occurs in causal inference where **V**1 and **V**3 are the potential outcomes under two treatments **V**1=1 or **V**1=0, where **V**2 is not observed on those receiving treatment **V**1=0 and **V**3 is not observed on subjects receiving treatment **V**1=1. The pattern of missing data could be exploited in the model specification or by breaking the estimation problem into simpler modular tasks. The another use of the pattern is to understand the limitation of the data or identify parameters that cannot be estimated. For instance from the above pattern (c) thereis a not on that no information to estimate the partial correlation between V2 and V3 conditional on V1.

3. Types of Missing Data:

In any clinical trials, data analysist or statistician or an expert must know the types of missing data and its mechanism is necessary. Suppose the number of cases of missing values is extremely small then an expert may drop or omit from the analysis. On the other side, if the missing values are huge and the volume of the data also morethen, it is not possible to omit or avoid, the impact will be more on the analysis data. Now, the type of missing data is to be discussed.

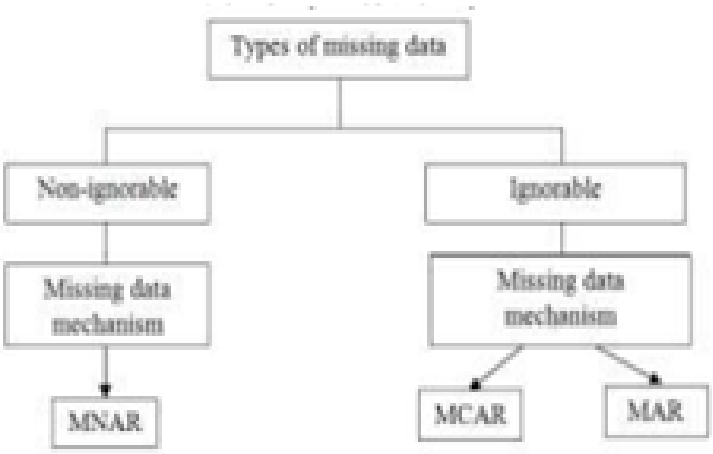

Figure 2: Types of Missing Data

In Figure 2, the types of missing data classified in a perfect manner. The main classification in Missing values are Non-Ignorable and Ignorable. Again from Non ignorable the data mechanism for Missing data is Missing not at Random (MNAR) only and in Ignorable missing data mechanism there are two different types are involved they are Missing completely at random(MCAR) and Missing at random(MAR). These of the different mechanism are applied in different situations to handling the missing values.Understanding the mechanism of handling missing data is important to get a unbiased and accurate decision results on the given

missing data. Missing data reduces the representativeness of the sample and can therefore affect the inferences about the population.

Most of the clinical trials, the experimenters can control the level of misingness and prevent the missing values before the data gathered. Missing value occurred due to the participant avoiding the questions in the questionnaire.In situations where missing values occurred, the clinical trial person is often advised to use the methods of data analysis that are robust to misingness. If an analysis is robust then there exists a mild or moderate violations of the technique's and produce a little bias or no bias or distortion in the conclusions drawn about the population.Now, the elaborations of these methods are below

Non Ignorable Missing Data:

As the name given, this is the method of assumption that the missing data mechanism is ignorable. That is the chance of avoiding these incomplete values for the further analysis. There is no need of an explicit model of the missing data process. In the true situation this Non ignorable mechanism will accounting for any measureable factors that govern the response probability.There are some typical situations for non-ignorable missing data such as

- If the essential variables that govern the missing data process are not available
- If there is a belief of responders differ from non-responders
- If the given data is with truncation

These types of Non ignorable missing data mechanism will be occurred in economics, sociology and Biomedical problems. In statistical aspect, the non-ignorable missing data involved with the problem of inference, estimation and model selection. In this Non ignorable missing data , the mechanism is Missing Not at Random(MNAR)

Missing Notat random (MNAR):

Missing Not At Randomis a type of missing data mechanism in Non-ignorable classification. It means that there is a relationship between the propensity of a value to be missing and its values. This is a case where theinvestigator with the lowest education is missing on education or the sickest people are most likely to drop out of the study. This missing not at random classification is non ignorable because the missing data itself to be modeled as deal with the missing data and thevalues that are missing will include some model with likely values. The missing data mechanism under non ignorable, if the failure to observe a value depends on the value that would have been observed other missing values in the given data. Missing not at random data are most common in longitudinal studies in which misingness is the outcome of the study dropout, illness and so on. In handling the Missing not at random mechanism, a valid inference requires specifying the correct model for the missing data mechanism, distributional assumptions for the response, or both. The resulting parameter estimates and statistics for testing of hypothesis may be a sensitive based on these assumptions. It is worth mentioning that unfortunately, one cannot determine whether misingness is MNAR or MAR solely based on the data at hand. For this reason, there is a growing consensus among statisticians studying missing data methods that a key component of an analysis is to carry out sensitivity analyses by fitting different missing data mechanisms to examine how sensitive the results are to the assumptions of whether misingness is missing not at random data.

Ignorable:

"Missing completely at Random" and "Missing at Random" are both considered 'ignorable' because we don't have to include any information about the missing data itself when we deal with the missing data.

Missing Completely at Random (MCAR):

Missing completely at Randomis one of the types that belong to Ignorable missing data mechanism. The other is missing at random mechanism. In missing completely at random, there is no relationship between the misingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is no systematic going on that makes some data more likely to be missing than others.

Data are said to be missing completely at random, if the failure to observe a value does not depend on any data, either observed or missing. Simple examples of this type of missing data mechanism include lost data, accidental omission of an answer on a questionnaire, accidental breaking of laboratory instrument, and personnel error. In an analysis of logistic regression, let us suppose that the response is completely observed for all participants, whereas some of the covariates are missing for some participants. Then the missing covariate values are missing completely at random if the chance of occurrence or simply probability of observing the missing covariate is independent of the response as well as independent of the values of the covariates that are fully observed or the covariates that would have been observed (ie, the missing covariates). Under this category of missing, the observed data are just a random sample of all the data. Bias is defined as the average difference between model parameter estimates and their true values.

Missing at Random (MAR):

Missing at Random**,** is another method of missing mechanism in ignorable category and it means there is a systematic relationship between the propensity of missing values and the Observed, but not the missing data. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the

values of an individual's observed variables. For instance, if men are more likely to tell you their weight than women, weight is missing at random mechanism.

Data are called as missing at random if, given the observed data, the failure to observe a value does not depend on the data that are unobserved. As an example, in cancer clinical trials, information on the size of a primary tumor is often missing, and the size of the primary tumor may depend on the type of the primary tumor, which is often fully observed. If the probability of primary tumor size is missing only based on the type of primary tumor, then the misingness is considered to be a missing at random technique. For a more common example involving missing covariates, suppose that the response is completely observed, whereas some covariates may be missing for some participants. The missing values of the covariates are missing at random, if, given the observed data, the probability of observing the missing covariate is independent of the values of the missing covariate that would have been observed, but this probability is not necessarily independent of the response or the fully observed covariates. Missing at random is a more realistic assumption than missing completely at random, but in this case, adjustments must be made, because the observed covariates are no longer a random sample. Clearly, if the missing data are missing completely at random, then they are missing at random. In most missing at random scenarios, an analysis will be both inefficient and biased. In data that are missing at random, if misingness depends only on the fully observed covariates and not on the response, then that analysis will lead to unbiased estimates. However, if the misingness based on the response variable (and not necessarily on the fully observed covariates), then a analysis will result in biased parameter estimates Missing data that are either missing at random or missing completely at random, along with the assumption that the parameters of the missing data mechanism are distinct from the parameters of the sampling model (ie, the joint distribution of the covariates and the response), are said to be ignorable missing. In these cases, the missing data mechanism can be ignored in making inferences about the parameters of the sampling model.

4. Statistical Methods for Missing Data in Clinical Trials:

After identifying or classifying the Missing data mechanism then by using several assumptions these missing data were handled with some analysis. Now it is the time to discuss the appropriate statistical method depends on the types of missing data mechanism that governs the misingness. In this section there will be a discussion on four most common methods for handling missing data and its statistical issues regarding missing outcomes versus covariates that are used in clinical trials. They are

Maximum likelihood (ML Methods)

A large class of model-based procedures arises from defining a model for the variables with missing values and making statistical inferences based on what are called maximum likelihood methods. Model-based methods are quite flexible and clearly set forth underlying model assumptions so that they can be evaluated. Additionally, standard errors can be easily obtained based on the model using appropriate algorithms and techniques, which take into account the missing data. The reviews on maximum likelihood methods for missing data are enormous and the literature is too many to list here. For instance, the literature on missing covariates alone is considerable

Multiple Imputations:

In statistics, imputation is the process of replacing missing data with substituted values. When substituting for a data point, it is known as "unit imputation". Multiple imputation (MI) has emerged as a popular technique for dealing with missing data problems. The technique of multiple imputations involves creating multiple complete data sets by filling in values for the missing data. Then, each filled-in data set is analyzed as if it were a complete data set. The inferences for the filled-in data sets are then combined into one result by averaging over the filled-in data sets. Many authors have described MI and some of its variants for missing covariates. It is important to mention here that the imputed values are random; therefore, this randomness must be captured in the standard errors of the parameter estimates. There are two types of multiple imputation techniques, which are called improper and proper imputation. Improper imputation uses an imputation model that is different from the analysis model, whereas in proper imputation, the imputation model is based on the analysis model. Improper multiple imputation yields biased estimators. Proper multiple imputation, although computationally more intensive, yields unbiased estimates with good large sample properties. Lastly, it is to be mention that the motivation and basis of proper multiple imputation is Bayesian, but the idea of multiple imputation itself is quite general and can be applied to other methods, such as hot deck imputation.

Fully Bayesian:

Fully Bayesian (FB) methods for missing data (covariate and/or response data) involve specifying distributions for all of the parameters usually called a prior distributions as well as specifying distributions for the missing data. The missing data are then imputed from these distributions. Bayesian methods can easily fit missing data without requiring new tools for statistical inference. In this sense, fully Bayesian methods are perhaps the most powerful and most general methods for dealing with missing data. Bayesian also connected with maximum likelihood and multiple imputations. The close connections between themaximum likelihood, multiple imputations and fully Bayesian procedures

Weighted Estimating Equations:

There are various other approaches requiring fewer model assumptions have been developed to account for missing observations. A common approach called weighted estimating equations (WEEs) has been proposed by Robins et al. .Weighted estimating equations methods are the missing data counterparts called generalized estimating equations That is GEE is a weighted estimating equations are generalized estimating equations methods adapted to the presence of missing data. General weighted estimating equationsare often called doubly robust in the sense that to obtain an unbiased estimate of the regression parameters, either the missing data mechanism or the estimating equations for the missing data given the observed data must be correctly specified, but not both. Weighted estimating equations methods for missing covariates were more important in missing values handling mechanism. It is good to mention here that GEEs are indeed a valid procedure when the missing data are missing completely at random, but they are generally biased if misingnessis missing at random or missing not at random.

Once a formal statistical model is formulated, estimation can be performed, by using any of the four formal methods mentioned, which is a far superior way to handle missing data compared with the ad-hoc complete case analysis method. These formal methods, however, can be computationally intensive and may not be available in standard statistical packages, such as SAS, STATA, or R. The maximum likelihood method is often viewed as the gold standard in model fitting. Maximum likelihoodmethods for generalized linear models are available in some statistical packages, and various versions of multiple imputations are available in SAS as well as other packages. Fully Bayesian methods are available in both SAS and Win BUGS, above discussed methods often produce similar results in many settings

Missing Outcomes Vs Covariates:

The impact of missing covariates versus that of missing outcomes on the estimates in a regression model may be quite different. First, is to be mention that if the data set only has missing responses and the missing responses are assumed to be missing at random, then a complete case analysis will lead to unbiased estimates of the regression coefficients. That means, a complete case analysis and amaximum likelihood analysis are equivalent. However, when the responses are missing not at random, then the complete case analysis as well as the maximum likelihood analysisassuming missing at random will lead to biased estimates. With missing covariates, the story is different. When we have only missing covariates in a regression analysis without missing responses, then the maximum likelihood analyses either assuming Missing at random or missing not at random are superior to the complete case analysis in terms of bias and efficiency of the parameter estimates. When we have both missing responses and covariates in a data set, as is common in longitudinal studies, then an maximum likelihood analysis either assuming missing at random or missing not at random is superior to the complete case analysis in terms of bias and efficiency as long as the sampling model and the missing data mechanism are assumed to be correct or approximately correct. If the sampling model and/or the missing data are badly misspecified in a maximum likelihood analysis, then the maximum likelihood method can yield bad results. Also mention here that in analyzing time-to-event data in a randomized clinical trial, one typically examines the treatment effect without adjusting for covariates. This is valid in cases where the censoring mechanism for the time to event only depends on the treatment group and does not depend on other covariates. When the censoring mechanism depends on other covariates, such as age, sex, and so on, then the assessment of the treatment effect should account for these covariates via a Cox regression in time-to-event studies. There are other simple adjustments in some specific scenarios like, using propensity score methods for adjusting for covariates in the assessment of treatment effects and the missing data indicator method to adjust for partially missing baseline measurements.

Conclusion:

Handling missing data is an important context not only in clinical trials but also in all trials. In any data the missing values used to affect the result of the conclusion and interpretations. It will produce a biased decision making about the data. To carry out such types of missing values in the given data, there is several methodologies available and it is to be based on the situation and criteria. In clinical trials handling missing data is necessary, yet difficult and complex task when analyzing results of randomized clinical trials. Considering the optimization technique for the handling of missing data during the planning stage of a randomized clinical trial and recommend analytical approaches which may prevent bias caused by unavoidable missing data.

After identifying or classifying the Missing data mechanism then by using several assumptions these missing data were handled with some analysis.Once a formal statistical model is formulated, estimation can be performed, by using any of the four formal methods mentioned, which is a far superior way to handle missing data compared with the ad-hoc complete case analysis method. These formal methods, however, can be computationally intensive and may not be available in standard statistical packages, such as SAS, STATA, or R. The impact of missing covariates versus that of missing outcomes on the estimates in a regression model may be quite different.There are other simple adjustments in some specific scenarios like, using propensity score methods for adjusting for covariates in the assessment of treatment effects and the missing data indicator method to adjust for partially missing baseline measurements

5. References:

1. Barnard, J.; Meng, X. L. (1999-03-01). "Applications of multiple Imputations in medical studies: from AIDS to NHANES". Statistical Methods in Medical Research. 8 (1): 17–36.
2. Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: John Wiley & Sons.
3. Pickles, A. (2005). Missing data, problems and solutions. In Encyclopedia of Social Measurement (pp. 689-694). Amsterdam: Elsevier.
4. Rubin, D. B. (1987). Multiple imputations for nonresponse in surveys. New York: John Wiley & Sons.
5. Rubin, D. B. (1996). Multiple imputations after 18+ years. Journal of the American Statistical Association, 91(434), 473-489.
6. Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: CRC Press.
7. Schafer, J. L. (1999). Multiple imputations: A primer. Statistical Methods in Medical Research, 8(1), 3-15.
8. Schafer, J. L., & Olsen, M. K. (1998). Multiple imputations for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33(4), 545-571.
9. Van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorial complex. British Journal of Mathematical and Statistical Psychology, 60, 315-337.