Title: Identifying Characteristics of Missing Data Mechanisms, Patterns, and Ratios: Implications for Statistical Analysis and Imputation Methods

Appendix 1:

The missing values mechanism concerns the relationship between missing-ness and the values of variables in the data matrix. So, this concept answers the difficult question of why the data are missing (12). Classifies the cause of the lack of data into three missing data mechanisms, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR when the occurrence of the missing data is independent of the data, MAR when the unavailability of the data depends on the values of observed variables, and MNAR when the process that causes the missing data depends on the values of the missing variable, and possibly observed ones too. While imputation depending on other covariates can be used for the first two types, subject-matter knowledge is required in dealing with the last type.

Missing values pattern defines which values are missing and observed in the data matrix, so this concept indicates where the missing data are. The missing values pattern is categorized into four groups uni/multivariate, monotone, general (or arbitrary), and file-matching pattern of missing data.

Given the significance of understanding the type of missing value mechanism in selecting the appropriate imputation method, we will provide further details on this topic.

Identifying the missing data mechanism is essential for selecting the appropriate method for valid analysis and minimizing biased results (26, 93, 94). If the data can be confirmed as (MCAR), a simple imputation method may be adequate. In the more common scenario of (MAR), multiple imputation is a beneficial approach (95). In longitudinal studies, a repeated measures model would be suitable. When the missing data mechanism is (MNAR), more advanced methods, such as joint modeling or pattern mixture models, should be utilized (24). Thus, comprehending the type of missing data mechanism is crucial for selecting the appropriate imputation method. Some methods for identifying the type of missing data mechanism are presented in Table 1.

Table 1: Different tests to identify the type of mechanism of missing values

missing values mechanism	MCAR	MAR	MNAR
Diagnostic methods for type of missing values mechanism	• Little's test (96) • Kim and Bentler test (97) • Jamshidian and Schott (98) suggest using homogeneity of covariances for testing MCAR • Fairclough's Logistic Regression Methods (For covariate and response variables) (12) • Jamshidian-Jalal-Non parametric (JJ-NP) test (for non-normal data based on testing Homogeneity of covariance) (96) • sensitivity analysis (99)	• Listing and Schlittgen (LS) test (100) (with pre-assumption is a monotone missing data pattern) • Fairclough's Logistic Regression Methods (For covariate and response variables)(12) • Nonparametric pair-wise variable (NVP) test (has good power, when missing data are due to truncation. For data that are either left truncated or right truncated, any of the methods involving means, such as Little-HM, KB-HM, or KB-HMC work well. The tests that involved testing Homogeneity of covariance (HC) such as KB-HC, JJ-H, and JJ-NP, generally had a better power to detect MAR than to detect MNAR.) (94) • sensitivity analysis (99)	• Comparison of immediate and reminder responders using Fairclough's method (Just for response variables) (12) • Nonparametric pair-wise variable (NVP) test (has very good power, when missing data are due to truncation. The power of these tests to detect MNAR increases with the correlation between the variables) (94)

Testing for MCAR is a logical initial step in the analysis of incomplete data. If the MCAR test is not rejected, several options for data analysis become viable. Conversely, if the MCAR test is rejected, the next question is whether the data are MAR or MNAR. As previously mentioned, if the data are MAR, multiple imputation methods within the category of traditional statistical methods are applicable. Testing for MNAR is not feasible without auxiliary information, as the missing-ness mechanism relies on the missing data itself (93). Additionally, the relationship between the missing data and the values of other variables can provide further insights into the underlying mechanisms of the missing data, which can be investigated using correlation analysis.

Figure 1 illustrates the process of using hypothesis testing to identify the type of mechanism for missing values, based on each of the diagnostic tests listed in Table 1.

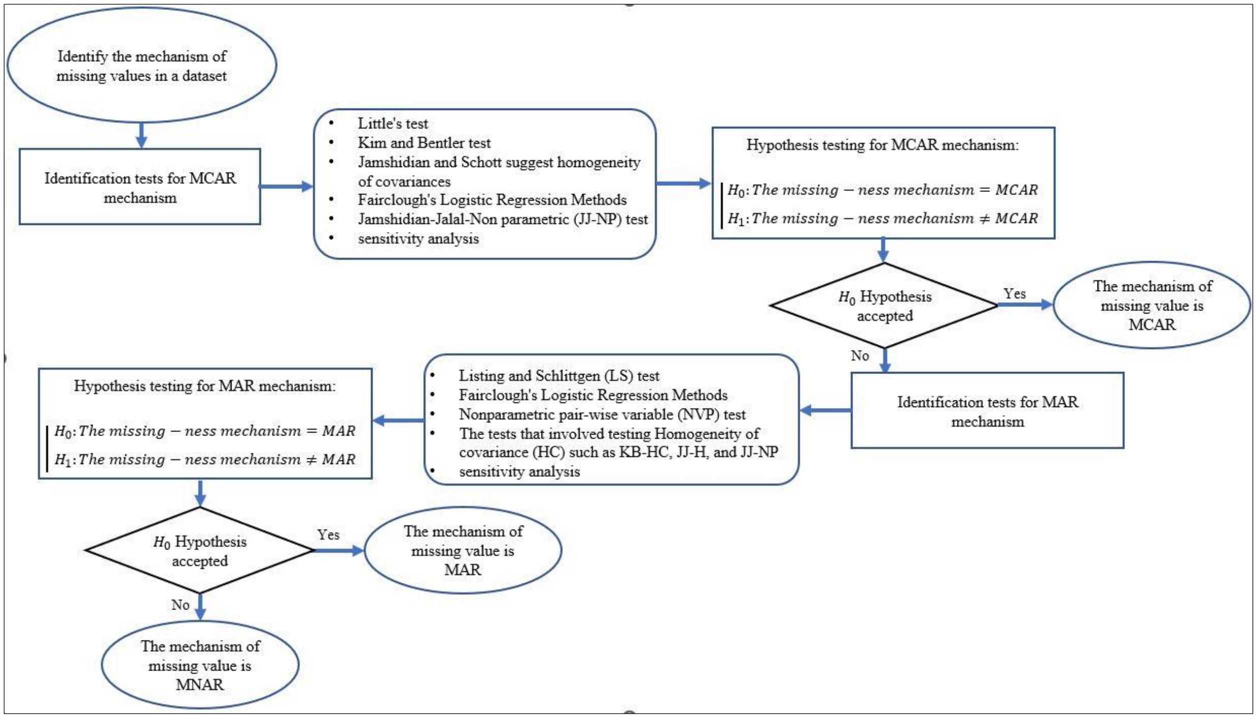

Fig. 1: Process of utilizing hypothesis testing to determine the mechanism behind missing values

The missing data pattern is crucial for selecting an appropriate imputation method to estimate missing values. Visualization tools should be utilized before imputation, (27). The md. pattern () function included in the multivariate imputation by chained equations (MICE) package helps explore missing data. Additionally, the Visualization and Imputation of Missing Values (VIM) package is highly effective for visually displaying these patterns. Three functions in the R programming environment—matrix plot(), ScattMiss(), and aggr()—can be employed for this purpose (27).

Identifying the subordinates of other concepts ratio of missing values in a dataset, study design, sample size, data type, the role of variables, distribution of variables, and correlation of variables, are easy to identify.

The ratio of missing-ness according to (101), the significance of missing values (MVs) in data can be categorized as follows: Less than 1% MV rate (Considered insignificant and can be generally ignored), (1-5)% MV rate (Considered trivial in real-world data sets and can be reasonably controlled), (5-15)% MV rate (Requires the use of very refined and sophisticated methods to effectively handle the missing-ness and impute the data accurately), More than 15% MV rate (Greatly impacts the predictive power and the reliability of inferences drawn from the data). In many real-world data sets, the missing value ratio often falls within the significant range, posing problematic challenges. Most standard statistical procedures are not inherently equipped to deal with the presence of missing values, which can lead to biased results and unreliable conclusions if not addressed properly. Therefore, it is crucial to carefully assess the extent of missing values in a given dataset and employ appropriate data imputation methods to handle the missing-ness, especially when the MV rates are high (5-15% or more). The choice of imputation method should be tailored to the specific characteristics of the data and the research objectives to ensure the integrity and validity of the analysis.

Study design this term refers to the comprehensive strategy or framework that guides the planning, execution, and analysis of a research project. It encompasses the methodologies utilized for data collection, measurement, and analysis, intending to effectively address specific research questions. Study designs can be categorized into two primary groups: observational study and experimental study.

The sample size is the number of individuals to be included in an investigation.

Data type this concept refers to a classification that specifies the type of value a variable can hold within a dataset. Data types determine the operations that can be performed on the data and influence its storage methods. Data types are primarily classified into two main categories: Qualitative or Categorical Data (This type of data cannot be measured or counted using numerical values) and Quantitative or Numerical Data (This type of data can be expressed in numerical values, rendering it countable and suitable for statistical data analysis).

Role of missing values this primary concept in the proposed framework addresses the occurrence of missing values within a dataset, specifically concerning independent (or predictor) variables, dependent (or outcome) variables, or both independent and outcome variables.

Distribution of Variables this concept refers to how a variable's values are spread or arranged across various categories or ranges.

Correlation of Variables this term refers to the interdependence between pairs of variables, describing the nature and strength of the relationship that exists between them.

All references used in Appendix 1 are identical to those cited in the main manuscript.