

## **IMPORTANCE OF DATA QUALITY**

Quality data is key to making accurate, informed decisions. And while all data has some level of “quality,” a variety of characteristics and factors determines the degree of data quality (high-quality versus low-quality). Furthermore, different data quality characteristics will likely be more important to various stakeholders across the organization.

### **1. Completeness**

Completeness is defined as a measure of the percentage of data that is missing within a dataset.

### **2. Timeliness**

Timeliness measures how up-to-date or antiquated the data is at any given moment.

### **3. Validity**

Validity refers to information that fails to follow specific company formats, rules, or processes.

### **4. Integrity**

Integrity of data refers to the level at which the information is reliable and trustworthy.

### **5. Uniqueness**

Uniqueness is a data quality characteristic most often associated with customer profiles. A single record can be all that separates your company from winning an e-commerce sale and beating the competition.

### **6. Consistency**

Consistency of data is most often associated with analytics. It ensures that the source of the information collection is capturing the correct data based on the unique objectives of the department or company.

## **DEALING WITH MISSING DATA OR INCOMPLETE DATA**

### **MISSING DATA**

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.

### **Types of Missing Data**

In general, there are three types of missing data according to the mechanisms of missingness.

- Missing completely at random
- Missing at random
- Missing not at random

### **Missing completely at random**

Missing completely at random (MCAR) is defined as when the probability that the data are missing is not related to either the specific value which is supposed to be obtained or the set of observed responses.

### **Missing at random**

Missing at random (MAR) is a more realistic assumption for the studies performed in the anesthetic field. Data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses, but is not related to the specific missing values which is expected to be obtained.

### **Missing not at random**

If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR).

The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may then be incorporated into a more complex one for estimating the missing values.

### **Techniques for Handling the Missing Data**

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully. The following are suggested to minimize the amount of missing data in the clinical research. One technique of handling the missing data is to use the data analysis methods which are robust to the problems caused by the missing data. An analysis method is considered robust to the missing data when there is confidence that mild to moderate violations of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population. However, it is not always possible to use such techniques. Therefore, a number of alternative ways of handling the missing data has been developed.

## **DATA CLASSIFICATION**

### **A Definition of Data Classification**

Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve. Data classification is of particular importance when it comes to risk management, compliance, and data security.

### **Types of Data Classification**

There are three main types of data classification that are considered industry standards:

- **Content**-based classification inspects and interprets files looking for sensitive information
- **Context**-based classification looks at application, location, or creator among other variables as indirect indicators of sensitive information
- **User**-based classification depends on a manual, end-user selection of each document. User-based classification relies on user knowledge and discretion at creation, edit, review, or dissemination to flag sensitive documents.