# **DATA QUALITY CHECKS PROCEDURES**

VA data consists of a lot of information gathered about a death that occurred at home. The data collection is done through mobile devices. In this process of data collection, errors can be encountered during data entry in the field. Therefore, performing data checks and data cleaning is an important procedure that must be done before analysis.

The quality checks that are performed on the data include:

1. For YES/NO kind of questions:

* Check for NAs or missing values in the data

1. For Dates

* Check for the range of dates
* Date formats
* Consistency of dates i.e. DOB vs DOD , DOD vs Date of registration

1. Gender

* Check for the consistency of allowed values i.e various coding schemes.

1. Age

* Check for negative values
* Check for extraordinary values i.e. ages above 120

1. Duration of symptoms

* Check for consistency of duration of illness vs duration indicated in the symptoms vs the age of the deceased.

1. Maternal pregnancy question vs duration of pregnancy vs age

* Check the consistency of age vs duration of pregnancy
* Also confirm the age, gender and maternal questions.

1. Consistency in age of respondents.

* Respondents should be at least of minimum age i.e. 15 years as required or specified by each country.

1. Relationship of respondent to the deceased.

* Check for consistency to avoid awkward cases of female fathers, male mothers, child of deceased neonate e.t.c.

1. Consistency of Age and level of education

* Check the consistency of age of deceased and the indicated level of education.

1. Duration of interview and the completeness of the VA.

* To ensure the VA was not a fake VA, check the duration of the full interview done (complete) with respect to the minimum time required for a VA interview. i.e a complete VA interview done in 10min is not logical.

1. Consistency of high level question with a low level one. These are questions that aks for general symptoms while another follow up question asks for a specific symptom. **e.g** **Id10228 (Did she have sores** and **Id10227 (Did she have sores or ulcers anywhere on the body)**
2. Confirm the consistency of coding schemes in the data and harmonize them e.g.

* Y,Yes,1 to represent YES
* m/M/1 for male OR f/2/0 for female

1. Duration of crying in minutes (check for values >30 and not 88/99)- remove 88/99 before analyzing COD

While exploring the variables in our data, we often see changes that need to be made.

Always make changes to a copy of the variable (or a copy of the entire data sets) **NEVER** change the original data.

Creating Yes/No, 0/1, or True/False indicators that flag a problem is also a good practice.

This process invloves 3 steps. . .

Find the “problem” (e.g., negative values for ages, people who were sick for 8,000 years, women with 14 children who are 20 years old)

Create an *index* that identifies the cases with the particular problem (or potential problem).

Create a copy of the variable and use the *index* to assign new values

**N/B:**

For each of the checks above, each country implementation should be able to discuss and document procedures to follow in solving them in case they exist and found in the data.

Data quality checks and data cleaning can be done using various tools that allow interaction with data. The commonly used tools that are open source and free include R-studio, Python and STATA. (There are also other tools).

After these checks, then the data will be ready for analysis and also to run Cause of death assignment algorithm on them (InterVA4, InterVA5, InsicoVA, Tarriff).