Chapter 13. Data Quality

This chapter discusses various aspects related to data quality.

13.1. Measuring data quality

Is the data complete? Is it collected on time? Is it correct? These are questions that needs to be asked when analysing data. Poor data quality can take many shapes; not just incorrect figures, but a lack of completeness, or the data being too old (for meaningful use).

13.2. Reasons for poor data quality

There are many potential reasons for poor quality data, including:

- Excessive amounts collected; too much data to be collected leads to less time to do it, and “shortcuts” to finish reporting
- Many manual steps; moving figures, summing up, etc. between different paper forms
- Unclear definitions; wrong interpretation of the fields to be filled out
- Lack of use of information: no incentive to improve quality
- Fragmentation of information systems; can lead to duplication of reporting

13.3. Improving data quality

Improving data quality is a long-term task, and many of the measures are organizational in nature. However, data quality should be an issue from the start of any implementation process, and there are some things that can be addressed at once, such as checks in DHIS2. Some important data quality improvement measures are:

- Changes in data collection forms, harmonization of forms
- Promote information use at local level, where data is collected
- Develop routines on checking data quality
- Include data quality in training
- Implement data quality checks in DHIS 2

13.4. Using DHIS 2 to improve data quality

DHIS to has several features that can help the work of improving data quality; validation during data entry to make sure data is captured on the right format and within a reasonable range, user-defined validation rules based on mathematical relationships between the data being captured (e.g. subtotals vs totals), outlier analysis functions, as well as reports on data coverage and completeness. More indirectly, several of the DHIS design principles contribute to improving data quality, such as the idea of harmonising data into one integrated data warehouse, supporting local level access to data and analysis tools, and by offering a wide range of tools for data analysis and dissemination. With more structured and harmonised data collection processes and with strengthened information use at all levels, the quality of data will improve. Here is an overview of the functionality more directly targeting data quality:

13.4.1. Data input validation

The most basic way of data quality check in DHIS 2 is to make sure that the data being captured is on the correct format. The DHIS 2 will give the users a message that the value entered is not on the correct format and will not save the value until it has been changed to an accepted value. E.g. text cannot be inputted in a numeric field. The different types of data values supported in DHIS 2 are explained in the user manual in the chapter on data elements.

13.4.2. Min and max ranges

To stop typing mistakes during data entry (e.g typing ‘1000’ instead of ‘100’) the DHIS 2 checks that the value being entered is within a reasonable range. This range is based on the previously collected data by the same health facility for the same data element, and consists of a minimum and a maximum value. As soon as a the users enters a value outside the user will be alerted that the value is not accepted. In order to calculate the reasonable ranges the system needs at least six months (periods) of data.

13.4.3. Validation rules

A validation rule is based on an expression which defines a relationship between a number of data elements. The expression has a left side and a right side and an operator which defines whether the former must be less than, equal to or greater than the latter. The expression forms a condition which should assert that certain logical criteria are met. For instance, a validation rule could assert that the total number of vaccines given to infants is less than or equal to the total number of infants.

The validation rules can be defined through the user interface and later be run to check the existing data. When running validation rules the user can specify the organisation units and periods to check data for, as running a check on all existing data will take a long time and might not be relevant either. When the checks are completed a report will be presented to the user with validation violations explaining which data values that need to be corrected.

The validation rules checks are also built into the data entry process so that when the user has completed a form the rules can be run to check the data in that form only, before closing the form.

13.4.4. Outlier analysis

The standard deviation based outlier analysis provides a mechanism for revealing values that are numerically distant from the rest of the data. Outliers can occur by chance, but they often indicate a measurement error or a heavy-tailed distribution (leading to very high numbers). In the former case one wishes to discard them while in the latter case one should be cautious in using tools or interpretations that assume a normal distribution. The analysis is based on the standard normal distribution.

13.4.5. Completeness and timeliness reports

Completeness reports will show how many data sets (forms) that have been submitted by organisation unit and period. You can use one of three different methods to calculate completeness; 1) based on completeness button in data entry, 2) based on a set of defined compulsory data elements, or 3) based on the total registered data values for a data set.

The completeness reports will also show which organisation units in an area that are reporting on time, and the percentage of timely reporting facilities in a given area. The timeliness calculation is based on a system setting called Days after period end to qualify for timely data submission.