CLEAR South Asia Centre | GEI Global Evaluation Initiative

Data Quality

1.1 Data Quality Dimensions

The effective use of administrative data rests on maintaining data quality. This entails maintaining accurate records and consistent processes on how the data are collected, collated, cleaned and cross-checked, independent data audits to ensure high accuracy, and ensuring safeguards in places where subjects have incentives to misreport information by implementing additional validation checks. Further, data quality standards can enhance operational and statistical uses of administrative data.

A high quality dataset should adhere to the following dimensions of data quality:

- **Accuracy** > How well does reflect reality?
- **Completeness** > Does the data fulfill the expectation of comprehensiveness?
- **Consistency** > Does the data stored in one place match relevant data stored elsewhere?
- **Timeless** > Is the data available as and when it is required?
- **Validity** > How accurately does the indicator map to the construct?
- **Uniqueness** > Is the only instance in which this data appears in the database?

A detailed description of the data quality dimensions is provided below:

1. Accuracy
Accuracy refers to the closeness of the administrative record data values to their (unknown) true values. It is important to check for information on any known sources of errors in the administrative data such as missing records, missing values of individual data items, misinterpretation of questions, and keying, coding, and duplication errors.

2. Completeness
Data completeness refers to the comprehensiveness of the data. For example, in the case of delivery of welfare programs, It is important to ensure that all the required data for identification and delivery of schemes and programmes is available in that dataset.

3. Consistency
Consistency refers to the overall reliability of the data. It is imperative to ensure that the dataset produces similar results under consistent conditions (such as when a type of analysis is repeated). Some errors to look out for while keeping a check on the consistency of the data are:
- Data collection errors: The procedure involved in primary data collection is prone to errors of mis-reporting, since the staff manually enters the respondents' information or response into the system. There could be errors in spellings, incorrect addresses, incomplete information etc.
- Incentives to mis-report: In the case that the performance of frontline workers is tied to outputs or outcomes such as school attendance, immunization completion, thresholds of malnourishment indicators, agriculture output etc, the staff in charge of service delivery may have an incentive to over-report or under-report outcomes based on the scheme benefits.

4. Timeliness
Timeliness refers to how well the administrative data meets the needs of the user at the time of need, as well as how up to date the data are.

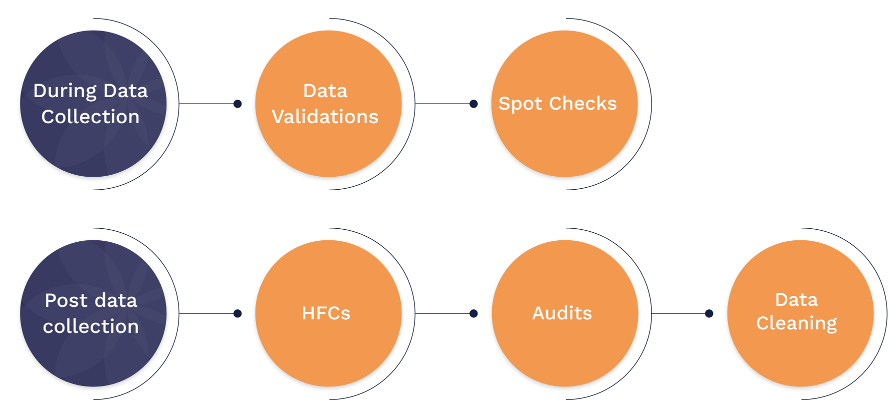
5. Validity
Validity is how accurately an indicator maps to the construct. A construct is a theoretical concept, theme, or idea which is not easy to measure. To measure this empirically, one needs to use "proxies" i.e. indicators. An indicator is "valid" if it can be measured accurately and without bias. For instance, cortisol levels may be an example of an unbiased measure of stress, where stress levels would be the construct But the stress test is still very noisy. While we think about the mapping of the indicator onto a construct, we would refer to this as the validity of the measure. Another example of validity (measures what it claims to measure) is a test of intelligence i.e. the construct, which should measure only intelligence and not something else (such as memory).

6. Uniqueness
Uniqueness of data means that there's only one instance of it appearing in a database. A common problem in databases is data duplication. In order to meet this data quality dimension it is important to review the information to ensure that none of it is duplicated.

1.2 Types of data quality checks

Inorder to ensure a high quality of administrative data being collected, it is imperative to incorporate quality checks at the time of data collection, as well as post-data collection. It is important to note that a significant amount of administrative data are collected using technology, where the data collection is either conducted by frontline workers actively or passively through machines. Researchers have used a variety of checks to ensure the data they collect is of good quality. The checks we describe below borrow from the checks conducted on survey data.

The type of data quality checks conducted for administrative data are mentioned in detail below:



Pre-data collection checks:

1. Data validations
The first iteration of data quality checks at the point of data collection can be in-built within the data collection application to reduce the possibility of human error. The different types of validations one can incorporate are as follows:

a. Soft constraints:
These constraints can be used to double check a response in cases where the responses for a particular variable can be unusual but not impossible. For example, if the input for a variable regarding the date of birth or Aadhaar card number are higher than a certain number of digits.
The following steps could be followed to incorporate a soft constraint in the form:

1: Identity scope for error
Analyse and select all the questions in the form which
have the possibility of having unusual responses.
Eg 1. Variables which are required for service delivery:
Please enter your Aadhaar card number.

2: Add extra question
Add a separate question (after each of those selected questions), confirming the inputted value. Add a relevance in the programming software, describing the condition in which the extra question should be asked.
Eg 1. You entered "xx" for your Aadhaar card number is this correct?
Relevance: Has to be exactly 12 digits

b. Hard constraints:
These constraints can be incorporated into a form to prevent responses that are impossible or extremely unlikely. For example, if age has been reported as less than zero. The following steps could be followed to incorporate a hard constraint in the form:

1: Identity scope for error
Analyse and select all the questions in the form which
have the possibility of having human input errors
Eg 1. Please input your age

2: Add a data validation
Add a validation in the programming software describing the constraint
Eg 1. The number cannot be <0 or >110

2. Spot Checks
Generally there are two ways in which data might be collected on field, a) the field team or implementation team might reach out to targeted individuals or households and directly collect data and input them using a digital platform or using paper based records e.g. ANMs collecting data at Anganwadis, personnel collecting data for a national level census survey, b) Another type of data collection happens where citizens or individuals may input of enter data using a digital platform (app or portal) e.g. for applying for citizen services eSewa, the citizens register or apply for a particular service.
Spot-checks are unanticipated visits by field staff (e.g. cluster resource coordinators or health visitors) to verify whether the data collection/ data recording is happening when and where it should be, at the time of service delivery. Such checks are particularly useful in the context of administrative data where there is an instance of active data collection.

The following things should be kept in mind when conducting spot checks:
a. Ensuring the appropriate usage of data collection tools are in place and are following the required compliance of the data collection processes.
b. Conducting spot checks on a higher percentage of processes at the initial stages of deployment of new hardware/software/data collection formats to catch errors early, and then to decrease the percentage checked over time.

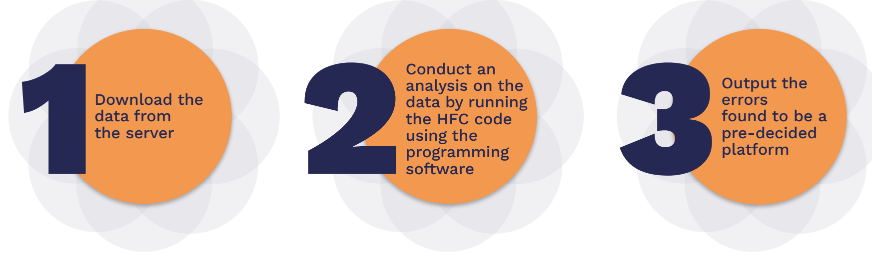
For example, in the context of administrative data collection, spot-checks can be particularly useful to check the data collection at points of service delivery such as provision of immunization services, Public Distribution Systems (PDS), etc.

Post-data collection checks:

1. High Frequency Checks (HFCs)
Inorder to ensure high quality of data, HFCs are routine checks made on incoming data, ideally on a daily basis, to check for data irregularities.

Objectives of conducting HFCs:
a. To keep a check on the accuracy of the data collected, for example by looking for information on any known sources of errors in the data.
b. To monitor the progress of data processes, and to measure the performance of data recording.
c. To check on the possibility of data fraud in administrative systems.

Before the data collection/data recording begins, a few steps should be taken in order to set up the HFC process:
a. Identify and create a list of the types of checks that would need to be conducted. A detailed list has been provided below, for reference.
b. Analyzing the data and creating dashboards to monitor these checks, This code would be run, in order to analyze the data every time the HFC is conducted.
c. Decide on how the errors that would be detected, after running the HFC, should be outputted. HFCs are run throughout the course of data collection. It is important to decide on the frequency of conducting the HFCs. HFCs should be run on a monthly basis, weekly basis, or every two days based on the size of the sample.



It is advisable to complete the setup of the HFCs before the data collection process begins. Once the data collection begins, the HFC should be run on a daily or bi-weekly basis.

The following steps can be followed at the time of conducting the HFC each time:

HFC Checklist

1. **Check for missing data**
a. Number and type of variables which have all missing values
b. % of missing observations in remaining variables

2. **Check for duplicates**
a. List of duplicate IDs which need to be rectified
b. Reasons for duplicates arising eg. Incorrect spelling, number written

3. **Check if there is any inaccuracy or outliers in the data**
a. If the values in some entries drastically higher or lower than the average entry value

4. **Check for the distribution of missing values and responses to skip-order questions**
a. If all valid skips have been compiled with and there are no invalid skips
b. If the entries are logically consistent (eg. age must increase with time)

5. **Completeness/validity of variables**
a. Validity - blank and other invalid entries eg. Phone numbers such as, '9999999999'
b. If the field implementers are not forging data eg. If they are entering their own phone numbers at the time of service delivery instead of the service receiver's phone number.

6. **Completeness/validity of variables**
If the data are comparable with similar data collection contexts (eg. the average house hold income in this data should not be drastically different from the average household income reported in another data source)

2. Backchecks or Data audits
In order to ensure that accurate data is being collected, it is imperative to set-up a system of field audits to verify the authenticity of collected information for a subset of records.

These audits should be carried out by a group of trained front line workers (independent to front line workers who collected the data originally), to revisit the place of data collection and verify the information which was collected.

The data collection format for the backcheck should be a shortened version of the original administrative data format which was used. Approximately 10-20% of the form entries should be randomly assigned to be backchecked.

While finalizing the shortened version of the administrative data form for the backcheck, the following variables that should included:

1. Type 1 variables: These are generally demographic questions such as about marital status, education level etc. which usually do not allow room for any error. Discrepancies between the original form entry and the back-check form entry indicates poor quality data.
2. Type 2 variables: These are variables which are key outcome variables for the program in question. Examples of such variables would be recall questions such as "Did the child get immunized" or "Did you go for an ANC visit?". Discrepancies between the original form entry and the back-check form entry indicates poor quality data.

The option of back-checks and audits may not be relevant for all types of administrative data. For example, in the case of income tax forms, the salaried income entered in the form can be corroborated with the income statements from the employers. However, in many cases backchecks and audits are relevant for the purpose of validating the same information from different sources to maintain data accuracy and for use as a tool for regulation of administrative processes using data.

The following steps can be followed to carry out an audit on a administrative data set:

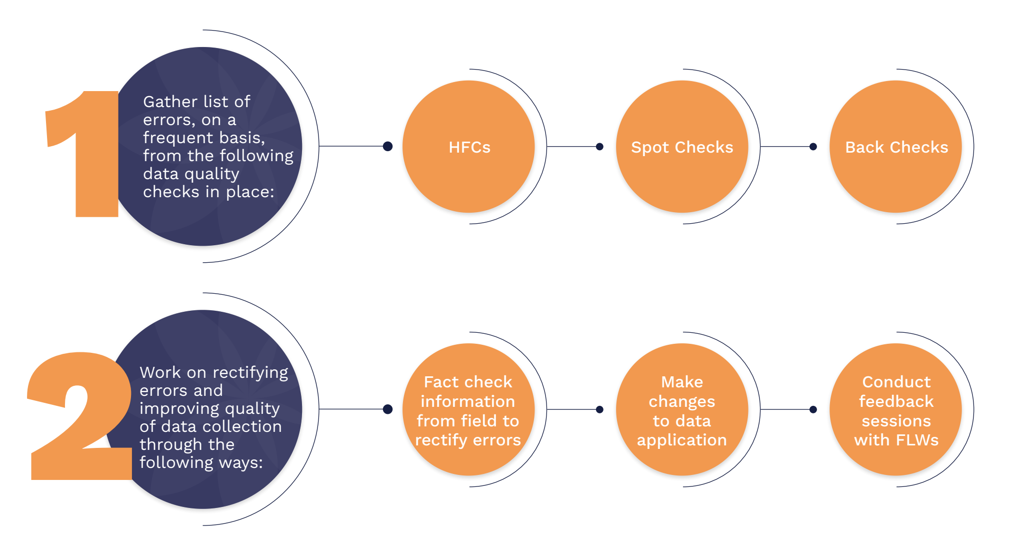
1. Sample a subset of the administrative data records that would be audited
2. Finalise a formal of the shortened version of the administrative data form
3. Finalise a group of trained FLWs (independent from the FLWs who collected the information originally), who would be carrying out the audits
4. The group of FLWs should visit the site, to conduct the data collection process with the shortened administrative data form (back check form)
5. The data should be reconciled and the rate of errors in the back check data with respect to the original data need to be calculated

Progress monitoring, feedback into processes, and data flow

Through this process of conducting timely spot-checks, HFCs and back-checks on the data being collected, an accurate list of errors in the data are identified.

As and when these errors are identified, there are various methods to rectify them while improving the overall quality of the data collection process:

Procure information from the field to be able to accurately correct the errors arising in the data.
Conduct feedback sessions with the front-line workers , to understand their challenges during data collection and explain to them the errors arising so as to reduce mistakes going forward.
Make relevant changes to the data collection application if that may prevent some common errors from arising going forward.



The above checks are primarily applicable for active administrative data collection. For passive data collection, only HFCs and subsequently making changes to the data application would be imperative.

Reference

• Data Quality Checks - JPAL Global
• Data Quality Assessment Tool for Administrative Data
• Data Quality Assurance Toolkit
• Checklist for Quality Evaluation of Administrative Data
• Data Quality Dimensions
• Data Quality Dimensions - ScienceDirect