REPUBLIKA Y'U RWANDA

UBUMWE - UMURIMO - GUKUNDA IGIHUGU

The Republic of Rwanda

nisr

NATIONAL INSTITUTE OF
STATISTICS OF RWANDA

Guidelines for Quality Assessment of Administrative Data

June 2018

National Institute of Statistics of Rwanda

Contents

Chapter 1: Introduction

The National Institute of Statistics of Rwanda (NISR) needs raw data for the production of official statistics. Apart from data obtained through surveys and censuses, NISR is increasingly using data collected and maintained by other institutions through their daily administrative records. The Administrative data is an example of such a data source (Wallgren and Wallgren, 2007). It is produced as a result of administrative processes of organizations but it is very often also an interesting data source for NSI's. During the last decade, more and more NSI's have realized this (Unece, 2007).

I.1. Definition of Administrative data

Administrative data refers to information collected primarily for administrative (not research) purposes. This type of information is routinely collected by Ministries, Government Departments and other organizations for the purposes of their normal business or operational activities, such as registration, transaction and record keeping, usually during the delivery of a service.

A major advantage of using administrative data for statistics compared to survey data is that it reduces the costs of data collection and reduces the administrative burden on enterprises and persons.

Since administrative data often covers whole populations, it is also very well suited for creating detailed and longitudinal statistics on subpopulations and regions (Wallgren and Wallgren, 2007). It has enormous potential to inform social science research and covers a wide variety of fields.

From a statistical point of view, administrative data also have some disadvantages. For example, the collection and processing of administrative data is beyond the control of the NSI. It is the data source keeper who manages these aspects, and not the NSI. The same is true for the units and variables an administrative data source contains. These are defined by administrative rules and may therefore not be identical to those required by an NSI (Wallgren and Wallgren, 2007). The disadvantages are predominantly the result of the fact that, in most cases, an NSI uses an administrative data source for a purpose different than the one for which the data was originally collected. As a result of this difference, the 'statistical' usability of a data source needs to be thoroughly studied by an NSI prior to its use.

Since NSI's want to produce high quality statistics (Statistics Netherlands, 2008), which are affected by the quality of the input data, it is of vital importance that NSI's are able to determine the quality of administrative data sources in an efficient and standardized way. For this purpose a quality framework developed by Statistics Netherlands enables the determination of the quality of secondary data sources, such as administrative data sources (Daas et al., 2008b).

As other NSI's, the National Institute of Statistics of Rwanda (NISR) use regularly data from different Ministries' departments, agencies and organizations to produce some indicators and intends to replace progressively some surveys by administrative data. This guideline will help to see if NISR collect really administrative data, if with the service providers they follow the standardised procedures for data quality assessment, in order to improve progressively the quality of this important source of information.

I. 2. Definition of Data Quality

Data quality is the perception of data's fitness to serve its purpose in a given context. Data are considered to be of high quality "if they are fit for their intended uses in operations, decision making, and planning" (Juran & Gryna, 1993). Additionally, the data are deemed of high quality if they correctly represent the real-world construct to which they refer. Data quality and data volume are generally negatively correlated in that as data volume increases, the issues of internal consistency within a database becomes more problematic, regardless of fitness for use for any external purpose.

For example, in longitudinal databases or a database where there are many records for an individual within a set time period, a person's gender, race, and birth date (DOB) can often differ between records. The more often information on a person is re-entered, the more likely the probability that differences will be found due to entry miscoding, different

People entering the information, and a host of other sources of potential inconsistency. Determining the accurate set of data elements becomes more difficult especially with high users of services.

Even though there is wide agreement on the need for good quality measures to assess the soundness of the data being used, most methods of assessing quality are generally ad hoc, informal and not very rigorous. The process of defining the way data quality is conceptualized and operationalized in an agency and the practice of writing down the methods being employed to estimate the size of uninsured populations, to present administrative information to policymakers, to help states assess the financial stability of their provider institutions as well to examine the health outcomes for the populations.

I. 3 Quality of Administrative Data Sources

The quality framework for administrative data sources is composed by several high level views on the quality of a data source, called categories (Batini and Scannapieco, 2006) or hyperdimensions (Karr et al., 2006). The latter term will be used in the remainder of this guideline. The quality aspects in each hyperdimension influence the usability of a data source in a different way. The following are three hyperdimensions are used to determine the Statistical usability of an administrative data source (Daas et al., 2008b):

- Source,
- Metadata, and
- Data

Each hyperdimension is composed of several dimensions; each dimension contains a number of quality indicators (**Table1**). A quality indicator is measured or estimated by one or more either qualitative or quantitative methods (Daas et al., 2008a-b). The Source, Metadata, and Data hyperdimensions each highlight different quality aspects of a data source. The hyperdimensions are also ordered according to an increasing level of detail. An important result of this ordered distinction is the fact that it efficiently guides the user in the study of the quality of a data source.

The purpose of this document is to guide the production of administrative data within the National Statistical System focusing on optimizing the quality. It starts with general concepts and considerations of administrative data, and it will show the status of administrative data in the context of Rwanda. The document can be used by any institution that keeps record through daily services, whose maintenance may generate administrative statistics.

Chapter 2: Hierarchical relation between the different aspects of quality used in the framework developed

II.1. Hyper dimension

II.1.1 Hyper dimension: Source

In the Source hyper dimension, the quality aspects related to the data source as a whole, the data source keeper and the delivery of the data source to the NSI are studied. The Source hyper dimension is composed of five quality dimensions:
Supplier; Relevance; Privacy and security; Delivery; and Procedures.

In Table1 the dimensions, quality indicators, and measurement methods for the Source hyper dimension are listed. In the Source hyper dimension mainly qualitative methods are present. Exceptions are the calculations of the effect of the use of the data source on i) the administrative burden induced by the NSI and on ii) the costs of the NSI (see **Table1**).

Table1: Dimensions, quality indicators, and methods for Source Hyper dimension and quality Indicators methods

Dimension	Quality indicator	Methods	
1. Supplier	1.1 Contact	Name of the data source	Name of the data source
	1.2 Purpose	Data source contact information	Data source contact information
		NSI contact person	NSI contact person
2. Relevance	2.1 Usefulness	Importance of data source for NSI	Reason for use of the data source by NSI
	2.2 Envisaged use	Potential statistical use of data source	Does the data source satisfy information demand?
	2.3 Information demand		Effect of data source use on response burden
	2.4 Response burden		
3. Privacy and security	3.1 Legal provision	Basis for existence of data source	Manner in which the data source is send to NSI; Are security measures required? (hard/software)
	3.2 Confidentiality	Does the Personal Data Protection Act apply?	
		Has use of data source been reported by NSI?	
	3.3 Security	Manner in which the data source is	

Dimension	Quality indicator	Methods	
		send to NSI	
		Are security measures required? (hard/software)	
4. Delivery	4.1 Costs	Costs of using the data source	
	4.2 Arrangements	Are the terms of delivery documented?	
	4.3 Punctuality	Frequency of deliveries	
		How punctual can the data source be delivered?	
		Rate at which exceptions are reported	
		Rate at which data is stored by data source keeper	
	4.4 Format	Formats in which the data can be delivered	

II.1.2 Hyper dimension: Metadata

The Metadata hyper dimension specifically focuses on the metadata related aspects of the data source. Clarity of the definitions and completeness of the Meta information are some of the quality aspects included. The Metadata hyper dimension is composed of four dimensions: Clarity, Comparability, Unique keys, and Data treatment (by the data source keeper). The Data treatment dimension is a special case. It consists of quality indicators used to determine whether the data source keeper performs any checks on and/or modifies the data in the source. This meta-information is very important for an NSI as it certainly affects the quality of the product delivered by the data source keeper (see **Table2**).

Table2. Dimensions, quality indicators, and methods for Metadata HYPERDIMENSION

Dimension	Quality indicators	Methods
1. Clarity	1.1 Population unit definition	Clarity score of the definition
	1.2 Classification variable definition	Clarity score of the definition
	1.3 Count variable definition	Clarity score of the definition
	1.4 Time dimensions	Clarity score of the definition
	1.5 Definition changes	Familiarity with occurred changes
2. Comparability	2.1 Population unit definition comparison	Comparability with NSI definition
	2.2 Classification variable definition comparison	Comparability with NSI definition
	2.3 Count variable definition comparison	Comparability with NSI definition
	2.4 Time differences	Comparability with NSI reporting periods
3. Unique keys	3.1 Identification	Presence of unique keys
		Comparability with unique keys used by NSI
	3.2 Unique combinations of variable	Presence of useful combinations of variables

4. Data treatment by data source keeper	4.1 Checks	Population unit checks performed
		Variable checks performed
		Combinations of variables checked
		Extreme value checks
	4.2 Modification	Familiarity with data modifications
		Are modified values marked and how?
		Familiarity with default values used

II.1.3 Hyper dimension Data

The Data hyper dimension focuses on the quality aspects of the data (facts) in the data source. Although the majority of the results focus on the quality aspects included in the Source and Metadata hyper dimensions, the Data hyper dimension is discussed here for completeness sake. The quality aspects of the Data hyper dimension are predominantly accuracy related with the exception of those included in the Technical Checks dimension. The quality indicators of this dimension can be seen in **Table3**.

Table3. Dimensions quality indicators methods of the data hyper dimension Quality Issues in the Use of Administrative Data Records

Dimension	Quality Indicator	Method
1. Technical checks	1.1 Readability	Can all the data in the source be accessed?
	1.2 Metadata compliance	Does the data comply with the metadata definition?
		If not, report the anomalies
2. Over coverage	2.1 Non-population unit	Percentage of units not

Dimension	Quality Indicator	Method
		belonging to population
3. Under coverage	3.1 Missing units	Percentage of units missing from the target population
	3.2 Selectivity	R-index 1) for unit composition
	3.3 Effect on average	Maximum bias of average for core variable
		Maximum RMSE 2) of average for core variable
4. Linkability	4.1 Linkable units	Percentage of units linked unambiguously
	4.2 Mismatches	Percentage of units incorrectly linked
	4.3 Selectivity	R-index for composition of units linked
	4.4 Effect on average	Maximum bias of average for core variable
		Maximum RMSE of average for core variable
5. Unit non response	5.1 Units without data	Percentage of units with all data missing
	5.2 Selectivity	R-index for unit composition
	5.3 Effect on average	Maximum bias of average for core variable
6. Item non response	6.1 Missing values	Percentage of cells with missing values
	6.2 selectivity	Index for variable composition
	Effect on average	Maximum bias of average for variable
		Maximum RMSE of average for variable
7. Measurement	7.1 External check	Has an audit or parallel test been performed?
	7.2 Incompatible records	Has the input procedure been tested?
	7.3 Measurement error	Size of the bias (relative measurement error)
		Fraction of fields with violated

Dimension	Quality Indicator	Method
		edit rules
8. Processing	8.1 Adjustments	Fraction of fields adjusted (edited)
	8.2 Imputation	Fraction of fields imputed
	8.3 Outliers	Fraction of fields corrected for outliers
9. Precision	9.1 Standard error	Mean square error for core variable
10. Sensitivity	10.1 Missing values	Total percentage of empty cells
	10.2 Selectivity	R-index for composition of totals
	10.3 Effect on total	Maximum bias of totals
		Maximum RMSE of totals

II.2 Application of the framework

II.2.1 Evaluation sequence

The framework introduced above is used for the determination of the quality of administrative and other secondary data sources. As a sequence, the user must first evaluate the quality indicators in the Source hyper dimension, then those in the Metadata hyper dimension, and finally those in the Data hyper dimension. This strict order approach is advised because it prevents that problems observed earlier on in the evaluation are (later on) found to be so severe that they block the use of the data source for the statistical application the user had in mind. When unsolvable problems occur during the evaluation of the Source hyper dimension it is likely that the user has to conclude that the data source cannot be used for statistics at all.

The quality aspects of the Metadata and Data hyper dimension should always be reevaluated for such data sources. If the evaluation of the last hyper dimension, Data, is successful, the data source can be used for the production of statistics. It is conceivable, however, that the user would like to perform one or more additional -very specific- checks after the evaluation of the three hyper dimensions (Kuijvenhoven and Schouten, 2008). These additional checks will occur at the data level.

II.2.2 Checklist

For the evaluation of the Source and Metadata hyper dimension, the authors who developed above mentioned framework have developed a checklist (Daas et al., 2008b; Daas et al., 2009). The checklist guides the user through the quality indicators that need to be evaluated for both Source and Metadata. For the Data hyper dimension a checklist cannot be used because of the large amount of calculations that need to be performed.

Since the predominant part of the methods in the Source and Metadata hyper dimension are qualitative, usually a score has to be filled in. When problems are found or a question cannot be answered completely, the user is guided in the steps to take. Apart from this, additional space is included to write down remarks.

Chapter 3: Quality Issues in the Use of Administrative Data Records

This Guideline provides an overview of “data quality” issues and the factors associated with them. These attributes will be applied to administrative data, with a focus on enrollment or eligibility and service claims data used in monitoring the cost and utilization of health care services. Issues of accuracy, comprehensiveness, and validity will be discussed and recommendations will be made for using administrative records given the current state of quality found in these data systems.

III.1 Requirements of administrative data use

Nowadays, administrative data systems are frequently required for different purposes:

- To record information and report on service use in agencies
- To evaluate programs for purposes of funding accountability and effectiveness of programs
- To measure the performance of for example the patient outcomes, and quality of care…
- To guide for national development plans and achievement evaluation.

The use of these data in decision-making requires personnel with particular expertise in data management evaluation design, and statistical skills.

The use of administrative data requires also the advancements in computing technology to enable providers/agencies to collect and report on service use in a cost efficient manner.

For example administrative data systems, like the American Medicaid and Medicare eligibility and service claims records, state and county event or encounter service data,…, and the like, are generally designed for internal reporting and reimbursement by a single agency or system.

But now states, as well as the federal government, are requiring service providers to send them administrative data from programs at an increasing level of detail for purposes of funding accountability and for monitoring the effectiveness of programs. These data are frequently used for measuring performance, patient outcomes, and quality of care (i.e. Healthcare Effectiveness Data and information Set, or HEDIS, measures).

The government of Rwanda through the National Statistics of Rwanda collects also data from different institutions, like MINEDUC, ITRS (BNR), RAB, REG, RRA, RBD, NIDA, RBC, HMIS data bases. Data collected help to calculate other indicators needed for police makers and other development agencies.

III.2 Challenges of Administrative data use

Major challenges of these data are the interpretability, coherence, and accuracy or quality of data items that are being integrated across programs and longitudinally over time.

Limitations and challenges exist also with respect to access or acquiring large administrative data files, data management, data integration, and, most importantly, data quality issues that are present at each step of the process.

Although there are many challenges using administrative data, there are also great opportunities:

- They are relatively inexpensive to use for evaluation, especially in longitudinal studies that track individual patients over time and across providers (Motheral & Fairman, 1997; Quam et al., 1993).
- They are a source of information on a large number of cases lending greater power for purposes of statistical inference (Motheral & Fairman, 1997; Garnick, Hendricks, & Comstock, 1994; Lohr, 1990). This makes them valuable in conducting population-based studies, detecting variations in practice patterns, and identifying specificities. Quality of care problems and health disparities that warrant further investigation (Iezzoni, 1997; Ballard & Duncan, 1994).
- They are beneficial in studying low prevalence disorders, such as schizophrenia, or rare events, where there is high service use and costs for a small percent of the population.

- When records have at least one similar personal identifier which is unique (e.g., social security number; first and last name), they can be readily linked and aggregated across organizations and systems to build a comprehensive-client level history that should be useful in treating patients with chronic comorbid conditions receiving treatment in different facilities or programs.
- They can also be used to monitor inappropriate drug utilization by clients and questionable provider prescription patterns
- Additionally, for Rwanda, this administrative data could allow to compute indicator at decentralize level

III.3 Data Diagnosis and Integration

Using secondary data for administrative, reporting, or research purposes entails multiple activities of which assessing and assuring quality is a major factor. There are, however, other essential activities that are interrelated with data quality issues that are integral to the process.

III.3.1 Data diagnosis

Data diagnosis involves initially assessing the data to understand its quality challenges. Data Profiling refers to inspecting data for errors, determining inconsistencies, checking for data redundancy, and completing partial or imperfect information. Profiling also includes a clear description of who the sample population is and the representativeness of the records to the universe that is being captured. Before any data set can be used, the number of records per quarter or year should be examined and compared to other administrative reports or other years of data.

If record or person numbers differ, further discussion is required to understand the source of the discrepancies. Additionally, frequency distributions of all variables should be examined to assess missing values, incorrect codes, outliers, etc. These records should be corrected or in some instances set aside in certain types of analyses. Duplicate records should be removed based on a predetermined set of criteria of what constitutes an exact replica. For example, numerous records for the same hospital stay are sometimes found in a data set based on billing practices of hospitals. These records, if not aggregated, can reflect multiple episodes for an individual. This problem sometimes occurs for residential treatment programs or other long term facility stays.

Data Integration is the process of matching, merging, and linking data for a wide variety of sources from disparate platforms. Matching or linking is a way to compare data so that similar, but slightly different, records can be aligned. Matching may use "fuzzy logic" to find duplicates in the data. For example, it often recognizes that 'Bob' and 'Robert' may be same individual

III.4. Data Augmentation and Monitoring

It might also find links between husband and wife or children at the same address. Finally, it can be useful in building a composite record, taking the best components from multiple data sources, and constructing a single super-record. For example, the most frequent name may be chosen as the true name when there are multiple records that should have "same" information for an individual. An example of linking data like mental health and substance abuse treatment for a single individual being treated in different systems is illustrated using a program known as "Link King". This program can be used for probabilistic links when all identifiers are not available and for deterministic links when identifying information is of good quality.

- **Data Augmentation** is the process of enhancing data information from internal and external data sources and involves the addition of any piece of related data*. Example s,*such as geocoding for name and address, can match data to US and Worldwide postal standards; phone numbers; contact information; common system wide identifiers associated with a case number at an agency, etc. all represent augmentation practices*.*
- **Data Monitoring** is making sure that data integrity is checked and controlled "overtime". Monitoring involves identifying variations in the data that require examination as to the cause. Software, based on certain algorithms, can be used to auto-correct variations if an error is involved. If the result is not inaccurate data, further exploration is required to understand changes in patterns due to policy or reimbursement or organizational changes which are not quality related.

For example, the Agency for Healthcare Research and Quality (AHRQ) and the Health Resources and Services Administration (HRSA) has an initiative to monitor the health care safety net (http://archive.ahrq.gov/data/safetynet/billings.htm).

The monitoring tool, “Tools for Monitoring the Health Care Safety *Net,”* aids administrators and policy makers in assessing local health care safety nets.

Chapter 4: Data Quality Components

IV.1. Accessibility

It refers to the availability of data in a warehouse and the ease of retrieval for monitoring, reporting, and analysis purposes. Accessibility refers also to the ease with which the data file extract can be obtained from the administrative agency. This includes the suitability of the form or medium for transferring the data, confidentiality constraints, and cost. There are always difficulties associated

Related challenges:

a. Internally, problems of sharing and confidentiality often prevent data from being used even between divisions of an agency that has different departments.
b. Technical problems associated with record retrieval and record transfer frequently occurs between divisions that sometimes serve the same clients through different programs.
c. Most agency data has not been prepared for external purposes, thus policies and procedures for data access by outside groups are unclear.
d. Some administrative data are generated by agencies that are not aware that they are producing data for statistical use. For that in some cases, key variable needed are not collected.

e. Procedures of accessing administrative data are complicated. These challenges apply to the Rwanda situation too.

IV.2. Data retrieval

Data retrieval means obtaining data from a database management system. Once access has been negotiated, other challenges exist.

Related challenges: Data retrieval has become more difficult with the use of complex data warehousing systems and staff that are often overworked or not permanent employees but contractors operating the databases.

IV.3. Security or Confidentiality

Sharing information can bring many benefits, and in the context of academic research, it is a time-efficient way of facilitating the dissemination of information. But sharing information also presents risks.

As information systems and the World Wide Web become more complex and widespread, so the potential for more information about individuals' private lives to become known without their consent or approval increases. When deciding to share personal information, the party seeking access and the data provider both need to identify the objective that it seeks to achieve, and then consider the potential benefits and risks of sharing the information. The privacy risk involved in any data sharing will depend on the specific circumstances of the data and the situation.

IV.4. Timeliness

Increasingly, the measures taken to protect data confidentiality create a delay before data sharing can occur, and can make the data less relevant for operational or policy decision making.

IV.5. Relevance

Relevance refers to how well the administrative data file meets the needs of the user in regards to data level (person, family, household, establishment, company, etc.), broad data definitions and concepts, population coverage, time period and timeliness.

IV.6. Interpretability

Interpretability refers to the clarity of information to ensure that the administrative data are utilized in an appropriate way. This includes evaluation of data collection forms, data collection instructions, and a data dictionary.

To be easily interpreted, administrative data producers must:

a. Describe each variable on the administrative data file and note their valid values
b. If a complete data dictionary is not available, describe the primary identification and response fields on the administrative data file and note their valid values.

IV.7. Target Population

A clear definition of who the individuals are and the percentage of that particular population that the data is collected on, must be provided.

IV.8. Accuracy/Coherence

Accuracy/coherenc**e is** related concepts pertaining to data quality. Accuracy refers to the comprehensiveness or extent of missing data, performance of error edits, and other quality assurance strategies.

- **Accuracy** is the closeness of results of observations to the true values or values accepted as being true. This implies that observations of most spatial phenomena are usually only considered to estimates of the true value. The difference between observed and true (or accepted as being true) values indicates the accuracy of the observations.
- **Coherence** is the degree to which data item value and meaning are consistent over time and comparable to similar variables from other routinely used data sources. Coherence of data and information reflects the degree to which the data and information from a single statistical program, and data brought together across data sets or statistical programs are logically connected and complete. Fully coherent data are logically consistent – internally, over time, and across products and programs.

IV.9. Data Records validity

Traditional reliability and validity assessments should be carried out on administrative files to establish utility of this information for performance or evaluation.

IV.10. Records- Comprehensiveness

A Comprehensive Data Definition refers to a formal data definition that provides a complete, meaningful, easily read, readily understood definition explaining the content and meaning of data. Data from administrative sources must be complete and easy to understand.

IV.11. Comparability of information collected at different time periods.

Administrative data collection is not ordered for evaluation purposes, thus a person may have their data collected (e.g. housing status, level of functioning) before or after an intervention of interest to an evaluator or policy maker. This often results in large variations in time between baseline information, intervention start up and follow up for subjects. Interpreting the results of analyses using administrative data is challenging especially when information is collected at different time periods and is not comparable for all subjects.

Chapter 5: Addressing Data Quality Challenges

This section presents different ways to assure data quality assessment. The main ways to deal with data quality challenges are:

V.1. Data management

The data Management is a broad term that refers to how data is structured or organized in a file, how it is stored (medium used), and what methods are used to protect i**t** (firewalls, backups, encryption)**.** Data quality is greatly affected by the way data is "managed," how accuracy is verified and consistency of information is addressed. The procedures and practices that support these processes must be well articulated and valued within an organization.

The good data management practices require:

- To verify the accuracy of collected data
- Regularly examine the information through diagnostic analysis
- Cleaning" the data that falls out of the boundaries and unduplicated records
- Assure that the data elements are standardized so that all data elements report the same item in the same way
- To give feedback reports to providers comparing their estimates with others in the system, as well as record reviews done on charts to check for similarities

- When there are discrepancies, other than minor ones, to signal the need to assess the input and output data processes to determine the source of the differences
- To have detailed data dictionaries
- To have data models
- To have information on how data and process flows within and between organizations
- Detailed specifications,
- To have regular internal and external audits and controls
- Encryption methods transform information in code
- Oversight: data steward, data custodian, or data management task council that oversees the data management decision-making process.
- Responsibility for the stewards to making sure that data elements have clear and unambiguous definitions, duplicates are eliminated, values are clearly enumerated or coded, and documentation is sufficient to allow suitable usage.
- To have solid data management procedures defining quality practices within or between agencies from the very beginning (to avoid the costs of correction and failed decision-making)
- Data elements must always be added, refined, or amended and reconstructed, particularly when used in longitudinal trend analyses, as the purpose of data changes over time to address new reporting and monitoring needs as well as new services.

V.2. Data storing and filing

Data storage refers to different ways of storing data. This includes hard copies, hard drives, flash memory, optical media, and temporary RAM storage.

When storing data, **confidentiality and data security** are important to consider.

Also **careful attention** and **continuous vigilance** when using administrative have **to be taken into account** for different reasons:

- As several records may exist for the same service (i.e., when payment claims are denied and then re-submitted), thus records must be unduplicated for analysis purposes.
- Multiple records for a single episode of inpatient when there are care that spans months may occur, therefore, adjustments must also be made
- Procedure codes can change over time, new services with a different name but similar function can create tracking difficulties in monitoring care.
- data storage and protection can be costly due to the confidential nature of these records

V.3. Verifying Data Accuracy

Management information staff must engage in multiple activities

- **externally** with those collecting and entering the data, **providers**,
- **internally** with those individuals storing and analyzing the data for planning and policy purposes, **NIS's.**

Providers should **be trained in data entry procedures** and should **have standardized definitions of all data items.**

Software applications should be developed that checks all data fields for formatting errors, field type and size, missing data, and checks for valid codes.

For example:

In Health facility, for example, when Health Management System only collect admission and discharge information for outpatient programs, missing records on client discharge or disenrollment from a program can lead to incorrect length of stay information.

Data submission reports can be generated for the data provider with total number of errors in each field, percent of accuracy and identification of outliers with questionable length of stay.

V.4. Data consistency

The following are some major approaches for verifying data consistency**:**

- To develop with the provider of data a corrective action plan by a specified deadline that details steps for correcting the data before their next submission
- To do retrospective audits
- To check all or a sample of the data against the original source

When the accuracy of the measure you are planning to use is found to be poor, consider dropping the variable or using a proxy measure in its place.

For example, use the history of substance abuse treatment as a proxy for co-morbidity if the records for drug and alcohol use are not identifiable due to confidentiality issues.

V.5. Verifying Data Consistency

Data cleaning practices is the essential practice for verifying data consistency.

This involves

- Eliminating duplicate records,
- Resolving differences in data elements among multiple records of the same individual or event in the same database (date of birth, gender, diagnosis, procedure),
- Resolving inconsistent data elements across databases when linking more than one data source (age, diagnosis, et al),
- Cross walking data elements over time within and between data sources, and constructing a new data element that is comparable within, between, and longitudinally over time.
- Rectifying the inconsistencies when there are different values for data elements, which should be similar
- Develop and implement Standardized rules to insure the integrity of the data elements that are immutable (date of birth, etc.) by:
  - ✓ Choosing the most frequent value,
  - ✓ Choosing the value found on the most recent record,
  - ✓ Constructing a variable in all the records that reflects this decision

- ✓ Or choosing the record or measure that is perceived to be most accurate based on its source of information (i.e. in reporting of demographic information found in a death certificate record)

V.6. Changing data items specification over time

Variation in data elements frequently occurs in monitoring events over multiple years within the same data source.

Changes in data systems often result in **new coding schemes** for the same variable. For example for this case, a variable such as case management may take on **different forms and meaning over time**, and the new data element may specify the inclusion of information that was formerly in a separate data element.

The change in the specification of the variable may require recoding or constructing a new variable for prior time periods.

Dealing with data consistency occurs again **when integrating information across systems**.

Thus, the variation in data element definition requires the "cross walking" of data elements if data sources are to be properly integrated.

This process requires clear definitions of the data element for each system and an ability to reconstruct data elements when the information differs.

Information generally needs to be aggregated to a higher or more general level when there are large discrepancies in the variable definitions

Information generally needs to be aggregated to a higher or more general level when there are large discrepancies in the variable definitions.

When the information collected differ for the same variables (e.i: date of birth/age; civil registration: birth certificate/health center certificate)

Chapter 6: Rwanda's Case

VI.1. Use of administrative data in Rwanda

In Rwanda there is two kind of institutions from which NISR or other institutions collect data: (1) Providers and (2) Producers.

Among providers we can notice:

- **Customs data**: Rwanda's customs data are collected by the use of software named Asycuda. The customs data have different quality issues related to valuation and quantification (recording quantity by use of proper supplementary unit). After NISR realized this issue, it has decided to continuously work with customs officials in order to make sure that these quality issues are addressed. This is done by regularly visit customs (at different borders) and sensitizes customs officials on how they should check these issues.
- **Migration Statistics:** Migration statistics (entry and exit) are collected by immigration officials stationed at different border posts. Migration statistics are used by NISR to gross up the Travel expenditure survey. These statistics may have quality issues related to proper classification of purposes of travel and country grouping. In addition to this, when there is a power shortage or an internet break up, it is difficult to capture the movement of persons (their numbers and purposes of travel). NISR has established a working

relationship with DGIE to make sure that some of these issues are dealt with through capacity building initiatives.

- **REG database:** The REG database can help provide records of imports and exports of electricity by country of origin (for imports) and country of destination (for exports). This can help the national accounts, BOP and trade statistics in general.
- **NAEB database:** NAEB database can help NISR to improve the quality of statistics of goods sold in auction either in Dar Es Salaam or Mombasa. This database can help with the exact date of the change of ownership, exact value of the transaction, and exact final destination.
- **RRA income tax:** This database, once it has data in the same detailed format, can help to know the declared annual income by company and by type of sector the company is involved in. This can help in identifying the biggest players in specific sectors and therefore facilitate the design of the sampling frame in different economic surveys.
- **RDB business registry:** This database can help us know the contact details of all registered businesses in Rwanda. It can easily facilitate in the sampling for enterprise surveys. Unfortunately, this database is not regularly updated. There is a need to put more efforts in having the regularly up-to-date business registry.
- **National Identification Agency** can help in building the updated National Population Registration database

and therefore facilitating different connected researches.

There are also institutions with proper data information system like:

- **International Transactions Recording system (ITRS):** This bank database managed by the National bank has all the banking transactions done by different economic actors. The purpose of ITRS is to make sure that all recorded transactions can be analyzed based on the purpose of transaction (whether the payment was made to buy a certain good or service). This can help the Balance of payments statistics to know the flow of money and its purpose. However, due to the fact that bank officials have limited knowledge on the usefulness of this data, a lot of work needs to be done to improve the quality of this database.
- **HMIS (Health Information system)/RBC:** The issue for this system is that the data collection tools do not allow to receive individual information, the data collected are global, like, number of deaths/health facilities, number of deaths, number of cases of malaria.

With this kind of data it is not easy to assess data quality, completeness and accuracy. NISR collect indicators already calculated with no information on the quality assessment.

There is no regulation document for formalizing collaboration between the HMIS and NISR for data quality assessment reporting

- The situation was the same, for the **EMIS** (**Education Information system**), but now NISR works with MINEDUC to develop a countrywide standardized data collection tool that enable to have information on each student, staff, and school infrastructure. A web based information system, is used to collect data but also to highlight missing and duplications and is progressively being improved. But there is no official regulation between NISR and MINEDUC for the quality assessment report.
- There is a project that implicates NISR, MOH/HMIS, MINALOC/NIDA, and sector officers in charge of Civil Registration and other vital events to improve the data collection of **CRVS** (Civil Registration and Vital statistics). The data collection tools and a web based application were developed for this purpose. But the completeness is not yet achieved.

VI.2. Opportunities for Rwanda administrative data collection

Based on the standards seen through this document, the following aspects can be observed. The government of Rwanda recognized the importance of administrative data as NISR and other agencies use it yet for statistical purpose. Even if the quality assessment does not yet meet the international standards, the opportunities exist for accelerating this quality assessment of administrative data.

- NISR has a department in charge of data quality issues including quality assessment of administrative data.

- There are different technical working groups grouping NISR as producers and departments as suppliers/providers'.
- Some organizations and institutions collecting data are working closely with NISR like MINEDUC, MINALOC / Civil Registration offices/NIDA, MOH/ RBC/ HMIS and started to improve the administrative data: data collection tools were developed, the web based application putted in place and the mechanism of data transfer known.
- Many agencies manifested their needs to improve the administrative data, and want technical support from NISR.

VI.3. Challenges related to administrative data

Although these opportunities, there are many problems that need to be resolved in order to improve administrative data:

- Lack of skills in assessing administrative data among NISR staff and NSS staff,
- Misunderstanding of the importance of collecting data by some external providers,
- Staffs that are afraid to lost their jobs which could be transferred to NISR,
- Risk of misreporting or over reporting due for the data used for evaluating 'imihigo',
- Still now, NISR collect indicators produced by agencies, without data quality assessment,
- There is no agreement between NISR and respective agencies describing the format of the data, the content,

the timing, the quality and the process of supplying data,

- The Quality assurance processes seen in this report are executed neither by provider nor by producers /users.

VI.4. What has to be done for quality assessment of Rwanda administrative data?

- Organize Conferences, meetings, technical workshops to sensitize NSS members on the importance of the administrative data, the good collaboration with NISR, on the necessity of putting in place a data quality internal audit and accepting an external audit from NISR,
- Organize practical technical trainings (hands on trainings) for NISR and NSS staff on the data quality assessment,
- Support NSS staff in charge of statistics to develop their respective data dictionary, and metadata
- NISR have to work closely with NSS member in order to negotiate the integration of the data needed in the NSS respective data,
- NISR have to work with NSS members to develop the Rwanda quality assessment framework,
- After training on data quality assessment, and the development of data quality framework, a Memorandum of Agreement (MOU) describing the format of the data, the population under study, the content, the timing, the quality and the process of supplying data, has to be signed,

- After the data quality framework is available, and the Memorandum of Agreement (MOU) signed, NISR have to start using the good procedures for data quality assessment by verifying first the Relevance, Interpretability, Comprehensiveness, validity, and Accuracy of data. Then NISR or other users could produce statistics if data are of good quality,
- NISR have to give regular support to the NSS members in charge of statistics for data quality assessment: harmonize data tools (data collection, metadata), define variables, execution of internal quality assessment.

Elaboration team

- Habimana Dominique
- Ntambara Juvenal
- Kambogo Francois
- Mukanyonga Apolline
- Mukarugomwa Jeanne d’Arc
- Uwayezu Beatrice
- Nzasingizimana Tharcisse
- Nyirimanzi Jean Claude

National Institute of Statistics of Rwanda
Po.Box 6139 Kigali Rwanda
www.statistics.gov.rw
info@statistics.gov.rw