FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA

MINISTRY OF HEALTH

HEALTH DATA QUALITY

TRAINING MODULE

PARTICIPANTS MANUAL

POLICY, PLANNING AND MONITORING & EVALUATION DIRECTORATE

JUNE 2018

FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA

MINISTRY OF HEALTH

HEALTH DATA QUALITY

TRAINING MODULE

PARTICIPANTS MANUAL

POLICY, PLANNING AND MONITORING &
EVALUATION DIRECTORATE

CONTENTS

Foreword V
Acknowledgements VI
List of Tables VII
List of Figures VII
Acronyms VIII
Section 1: Introduction 2
1.1 Module Description 2
1.2 Module goals 2
1.3 Module learning objectives 2
1.4 Description of training methods 3
1.5 Target Group 3
1.6 Core Competencies 3
1.7 Module duration and class size 3
Section 2: Introduction to Data Quality 5
2.1. Data and Data Quality Definitions 6
2.2. Importance of data quality 6
2.3. Leadership in data quality 7
2.4 Symptoms of data quality problem 8
2.5. Challenges in overcoming problems related to data quality 9
2.6 Possible solutions to problems of data quality 9
Section 3: Health Data Quality Dimensions 11
3.1 Introduction to data quality dimensions 12
3.2 Definitions and examples of the data quality dimensions 12
Section 4: Data Quality Assurance 26
4.1. Quality Assurance and Data Quality Assurance 27
4.2. Techniques of data quality assurance 27
4.2.1 Data Quality Desk review 31
4.2.2. Lot Quality Assurance Sampling (LQAS) 32
4.2.3. Visual Scanning (Eye Balling) 36
4.2.4. Routine Data Quality Assessment (RDQA) 36
Section 5: Using DHIS2 to improve data quality 58
5.1. Section Introduction 58
5.2. Data input validation 58
5.3. Min and max ranges 59
5.4. Validation rules 59
5.5. Outlier analysis 59
5.6. Completeness and timeliness reports 59
References 60
Annexes 61

FOREWORD

The Federal Ministry of Health is currently implementing the Health Sector Transformation Plan (HSTP), a five year strategic plan from 2015/16-2020. Information Revolution is one of the four transformation agendas of HSTP with the objective of maximizing the availability, accessibility, quality, and use of health information for decision making processes through the appropriate use of ICTs to positively impact the access, quality, and equity of healthcare delivery at all levels.

Improving data quality and promoting the culture of information use is at the center of the information revolution agenda. As a result, the Policy, planning and Monitoring &Evaluation Directorate (PPMED) of the FMOH has developed this data quality training manual which can be helpful for health workers and managers at all levels of the health system. It will be a useful guide to improve health data quality and measure data quality at health centers, hospitals, woreda Health Offices, Zonal Health Departments, Regional Health Bureaus and other health institutions

I would like to thank Monitoring and Evaluation Case team experts at PPMED, experts from RHBs, Universities and partner organizations for their great contribution in the finalization of this manual.

Biruk Abate

Director of Policy, Planning and Monitoring & Evaluation Directorate,

Federal ministry of Health

ACKNOWLEDGEMENTS

This manual is developed with inputs from a number of experts from FMOH, RHBs, Universities and partner organizations. The Policy, Planning and Monitoring & Evaluation Directorate of the Federal Ministry of Health is grateful for all who have been involved in the preparation of this manual.

LIST OF TABLES

Table 1: Internal consistency outliers 14
Table 2: Example of outliers in a given year for a certain indicator. 15
Table 3: Internal consistency: Trends over time 16
Table 4: Example of trends over time 16
Table 5: Internal Consistency: Comparing selected Related Indicators 17
Table 6: Example: Internal Consistency 17
Table 7: External Consistency: Compare with Survey Results 18
Table 8: Example: External Consistency 19
Table 9: External Comparison of Population Data 20
Table 10: External Comparisons of Population Denominators 20
Table x: Frequency of data quality techniques applied by administrative unit level and health facility 55
Table XX: Data quality techniques applied and data quality dimension addressed by administrative unit level and health facility 56

LIST OF FIGURES

Figure 1 Roles and Responsibilities of each level of the health system for maintaining data quality (From Measure Evaluation) 8
Figure 2 Major Differences among LQAS, DQA, RDQA and PRISM 29
Figure 3 Ethiopian Data Quality Assurance Timeline 30

ACRONYMS

AHU	Administrative Health Unit
ANC	Antenatal Care
CPD	Continuous Professional Development
DHIS	Distinct Health Information System
DHS	Demographic Health Survey
DQA	Data Quality Audit
DV	Data Verification
EMR	Electronic Medical Record
FMoH	Federal Ministry of Health
HC	Health Centre
HCWs	Health Care Workers
HF	Health Facility
HIS	Health Information System
HIV	Human Immunodeficiency Virus
HMIS	Health Management Information System
HP	Health Post
ICD	International Classification of Disease
IUCD	Intra Uterine Contraceptive Device
LQAS	Lots Quality Assurance Sampling
M&E	Monitoring & Evaluation
NCoD	Non-Communicable Disease
ODF	Open Deification Free
OPD	Out Patient Department
PMT	Performance Monitoring Team
PRISM	Performance of Routine Information System Management
RDQA	Routine Data Quality Assurance
RHB	Regional Health Bureau
RHIS	Routine Health Information System
SBA	Skill Birth Attendant
SNNPR	Southern Nations, Nationalities and People Region
TB	Tuberculosis
ToT	Training of Trainers
VF	Verification Factor
WorHo	Woreda Health Office
WHO	World Health Organization
ZHD	Zonal Health Department

SECTION 1: INTRODUCTION

1.1 MODULE DESCRIPTION

High-quality data are at the core of program activities. Availability of quality data is at the heart of a functioning evidence-based decision making in the health sector. It is widely recognized that quality data leads to better clinical and health admin decisions that results in better health outcomes for the country.

The Federal Ministry of Health (FMoH) has been working towards continuously improving data and information quality within the health sector. The Ministry reformed the health management information system in 2008 with the objective of ensuring improved measurement and standardization towards improvement in quality of data – enabling better decisions and thus better health outcomes. The reform registered significant improvements in availability and completeness of source documents and report accuracy. However, data quality is not at the required level and a lot has to be done if the data is to be relied upon to inform decisions on health policy, health programs, and allocation of resources.

1.2 MODULE GOALS

The overall goal of this training modules to improve data quality at all levels in the health system, by upgrading knowledge, skills, and attitude of health care workers, health information managers, and administrators at all levels on techniques of improving quality of health care data in all its dimensions. The module is designed to address all areas in health care where data are collected and information generated.

1.3 MODULE LEARNING OBJECTIVES

By the end of this module, participants will be able to:

- Identify the main causes of poor data quality
- Explain different dimensions of data quality
- Identify the roles and responsibilities of the different levels in the health system for maintaining data quality
- Define, calculate, and interpret data-quality metrics
- Differentiate the commonly used tools and methods for assessing data quality
- Define and describe the value of monitoring and using data-quality assessment results over time

1.4 DESCRIPTION OF TRAINING METHODS

- Interactive Lectures
- Group discussion and presentation
- Activity-based site visit
- Case studies

1.5 TARGET GROUP

- Health care workers
- Health Extension Workers
- Health Administrators from Woreda to Federal levels
- Academia

1.6 CORE COMPETENCIES

By the end of the training, the participants should be able to conduct the following tasks:

- Maintain data quality standards
- Able to use data quality assessment techniques
- Develop action plan for improvement of data quality

1.7 MODULE DURATION AND CLASS SIZE

- The module will take a total of five training days
- The maximum number of trainees for this module should not be more than 30.

SECTION 2: INTRODUCTION TO DATA QUALITY

Duration: 2 hours

Section Objectives

At the end of this Section, participants will be able to:

- Describe the concepts of data quality and its importance
- Identify symptoms of data quality problems
- Discuss the role of leadership in data quality management
- Explain the roles and responsibilities of each level of health system for maintaining data quality
- Discuss the potential challenges and possible solutions of data quality

Teaching Methods

- Brainstorming
- Interactive Lecture

Materials Needed

- Flipchart
- Tape
- Markers
- PowerPoint presentation
- Projector

Section Activities:

Activity duration: 30 minutes

Activity: Discuss what data quality means

- Actively participate on the brainstorming Section in small groups
- Write your responses on flipchart
- Compare your responses with the standard definition of data quality

2.1. DATA AND DATA QUALITY DEFINITIONS

Data is a key ingredient to improving health care quality. It is starting point for health care information, whether maintained manually or electronically at a large teaching hospital, health center or health post. Demographic and clinical data stored in a patient's medical/health record as well as the family folders are the major source of health information in Ethiopia.

What is data quality?

Data quality is often defined as **"fitness for use."**

What does this mean?

- Data are fit for their intended uses in operations, decision making, and planning.
- Data reflect real value or true performance.
- Data meet reasonable standards when checked against criteria for quality.

In general terms, quality data represent what was intended or defined by their official source, are objective, unbiased and comply with known standards.

2.2. IMPORTANCE OF DATA QUALITY

Good quality health is dependent on the access to and use of good quality data. The importance of good quality data includes:

For patient/Client

- Service users are more likely to receive better and safer care if healthcare professionals have access to accurate and reliable data to support decision making. Accurate and reliable patient data, such as results of investigations, information on allergies, past medical history, potential drug interactions, when readily accessible to the healthcare professionals supports provision of quality healthcare services.
- Service users are more likely to receive better care if healthcare performance data used to support quality improvement is of good quality and reflects actual performance.

For Healthcare organizations

- Healthcare organizations institute quality improvement initiatives based on performance measurement.
- Healthcare organizations can more effectively and efficiently plan and provide for service user needs if the data used to support decision making is of high quality. For example, good quality

demographic data that highlights an aging population or a significant increase in immigrants in a specific catchment area can enable organizations plan for the specific needs of that area

For Researchers

- Healthcare research contributes to improved outcomes by providing evidence to support particular care processes. This research can only be relied on if it is based on good quality data.

2.3. LEADERSHIP IN DATA QUALITY

Leadership can be defined as the process in which one engages others to set and achieve a common goal, often an organizationally defined goal (Robbins & Judge, 2001). Leadership in data quality management is the high-level policies and strategies that define the purpose for collecting data, the ownership of data, and the intended use of data. Leaders in data quality management are expected to ensure that health data is compliant with regulation, standards, and organizational policies.

Many health care administrators already recognize that quality improvement is the way to add value to the services offered and that the dissemination of quality data is the only way to demonstrate that value to health care authorities and the community. To ensure better quality health data all health workers and managers at each level should convey their role and responsibilities.

Activity: In your group discuss on the following points:

- Discuss the role of health workers and managers to ensure data quality and categorize by level (HF, intermediate administrative level (ZHD/WorHo) and central level (RHB/FMoH).
    - Actively participate on the brainstorming session in small groups
    - Raise your discussion points to the whole class

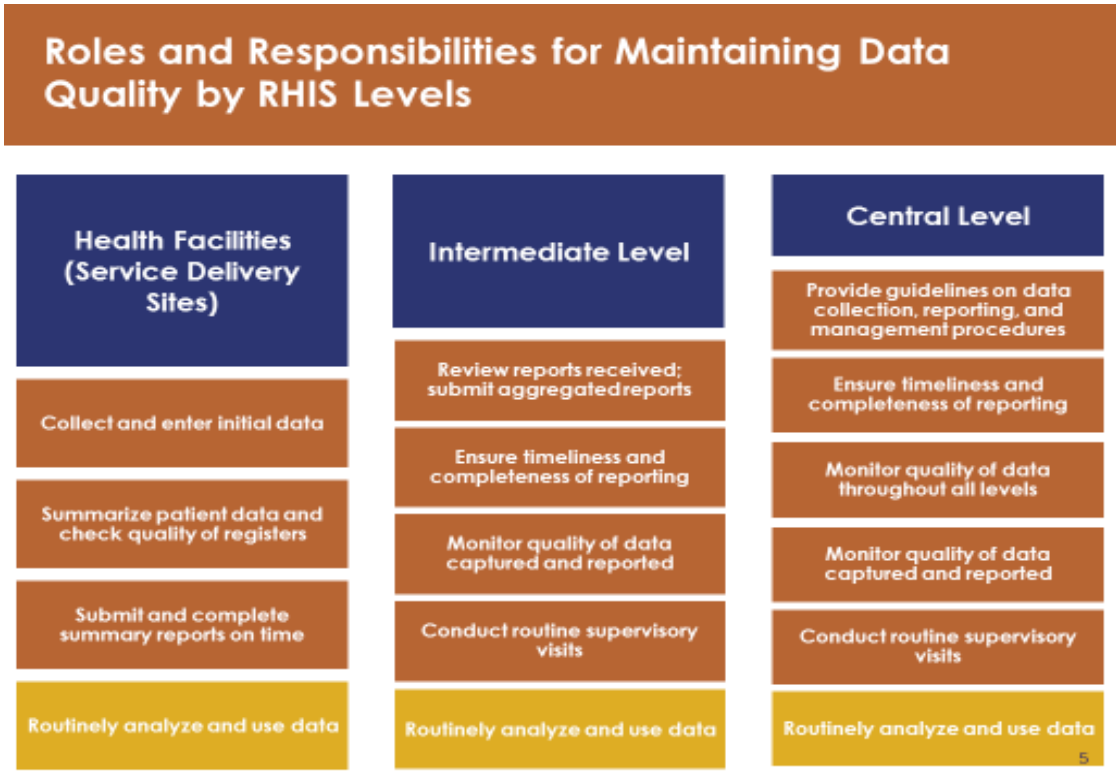


Figure 1 Roles and Responsibilities of each level of the health system for maintaining data quality (From Measure Evaluation)

Activity: Discuss the symptoms of poor data quality

- Actively participate on the brainstorming session in small groups
- Write your responses on flipchart
- Compare your responses with the identified symptoms of data quality problems

2.4 SYMPTOMS OF DATA QUALITY PROBLEM

- Different people supply different answers to the same question.
- Data are not collected in a standardized way or objectively measured.
- Staff suspect that the information is unreliable, but they have no way of proving it.
- There are parallel data systems to collect the same indicator.
- Data management operational processes are not documented.
- Data collection and reporting tools are not standardized; different groups have their own formats.
- Too many resources (money, time, and effort) are allocated to investigate and correct faults after the fact.
- Mistakes are spotted by external stakeholders (during audits).

2.5. CHALLENGES IN OVERCOMING PROBLEMS RELATED TO DATA QUALITY

Group Discussion-1

Activity: In Your small group identify the five most common problems you think that affect the quality of data and propose actions that could lead to improvements in data quality

Data quality can be affected by different problems across system level some of them are the following:

Technical determinants

- Lack of guidelines to fill out the data sources and reporting forms
- Data collection and reporting forms are not standardized
- Complex design of data collection and reporting tools

Behavioral determinants

- Personnel not trained in the use of data sources & reporting forms
- Misunderstanding of how to compile data, use tally sheets, and prepare reports
- Math errors occur during data consolidation from data sources, affecting report preparation

Organizational determinants

- Lack of a reviewing process, before report submission to next level
- Organization incentivizes reporting high performance
- Absence of culture of information use

2.6 POSSIBLE SOLUTIONS TO PROBLEMS OF DATA QUALITY

- Standardization and simplification of guidelines, and recording and reporting formats across the health system.
- Integration and institutionalization of health data
- Build capacity of health work force from data generation to information use
- Staffing of health institutions with necessary skilled human power to support the HIS
- Strengthen the Performance Monitoring Team (PMT) at each level of the health system
- Enhance culture of information use at each level of health system

SECTION 3: HEALTH DATA QUALITY DIMENSIONS

Section duration: 4:30 hours

Teaching Methods

- Interactive lecture
- Group discussion
- Group presentation
- Case study

Materials Needed

- Power Point presentations
- LCD Projector
- Flip charts
- Markers

Objectives

At the end of this Section, participants will be able to:

- Describe the different dimensions of data quality
- Explain how the data quality dimensions measured

Activity: what are the different data quality dimensions?

- Actively participate on the brainstorming session in small groups
- Write your responses on flipchart
- Compare your responses with the standard data quality dimensions.

3.1 INTRODUCTION TO DATA QUALITY DIMENSIONS

Regardless of whether in a hospital, health center, a clinic, or a health post, the quality of health care data and statistical reports has come under intensive scrutiny in recent years. Thus, all health care service providers, including clerical staff, health professionals, administrators, and health information managers, need to gain a thorough knowledge and understanding of the key components of data quality and the requirements for continuous data improvement.

Major dimensions of data quality are:

Dimension 1: Accuracy and Validity

Dimension 2: Consistency

Dimension 3: Completeness

Dimension 4: Timeliness

Sub dimensions of data quality:

Dimension 5: Legibility:

Dimension 6: Accessibility

Dimension 7: Confidentiality

Dimension 8: Precision

Dimension 9: Integrity

Dimension 10: Relevance

3.2 DEFINITIONS AND EXAMPLES OF THE DATA QUALITY DIMENSIONS

1. Accuracy and Validity

Accurate data are considered correct: the data measure what they are intended to measure. Accurate data minimize error (e.g., recording or interviewer bias, transcription error, sampling error) to a point of being negligible.

The original data must be accurate in order to be useful. If data are not accurate, then wrong impressions and information are being conveyed to the user. Documentation should reflect the event as it actually happened. Recording data is subject to human error and steps must be taken to ensure that errors do not occur or, if they do occur, are picked up immediately.

Question!

Give examples on accuracy and validity in both manual and electronic record system.

Example of accuracy and validity in a manual medical record system

- The patient's identification details are correct and uniquely identify the patient.
- All relevant facts pertaining to the episode of care are accurately recorded.
- All patient/client records (Cards, forms) in the integrated individual folder are for the same patient.
- The patient's address on the record is what the patient says it is.
- Documentation of clinical services in a hospital or health center is of an acceptable predetermined value.
- The vital signs are what were originally recorded and are within acceptable value parameters, which have been predetermined and the entry meets this value.
- The abstracted data for indices, statistics and registries meet national and international standards and have been verified for accuracy.

In a manual system, processes need to be in place to monitor data entry and collection to ensure quality. In a computerized system, the software can be programmed to check specific fields for validity and alert the user to a potential data collection error. Computer systems have in-built checks such as edit and validation checks, which are developed to ensure that the data added to the record are valid. Edits or rules should be developed for data format and reasonableness, entailing conditions that must be satisfied for the data to be added to the database, along with a message that will be displayed if the data entry does not satisfy the condition. In some instances, the computer does not allow an entry to be added if it fails the edit. In other instances, a warning is provided for the data entry operator to verify the accuracy of the information before entry.

Examples of edits and validity in a computer-based system

- In an electronic medical record (EMR) system, a patient must have a unique number because it is the key indexing or sorting field.
- The patient's number must fall within a certain range of numbers or the computer does not allow the data entry operator to move to the next field or to save the data.
- For hospital or health center patients, the date of admission must be the same as or earlier than the date of discharge.
- A laboratory value must fall within a certain range of numbers or a validity check must be carried out.

- Format requirements such as the use of hyphens, dashes or leading zeros must be followed.

- Consistency edits can be developed to compare fields – for example a male patient cannot receive a pregnancy test.

2. Reliability (Consistency)

Data should yield the same results on repeated collection, processing, storing and display of information. In other words, data should be consistent.

Dimension 2.1: Internal consistency of reported data

Internal consistency of the data relates to the coherence of the data being evaluated. Internal consistency metrics examine: 1) coherence between the same data items at different points in time, 2) coherence between related data items, and 3) comparison of data in source documents and in national databases.

Four metrics of internal consistency are included in the DQR. These are:

1. **Presence of outliers:**
2. **Consistency over time:**
3. **Consistency between indicators:**
4. **Consistency of reported data and original records:**

Dimension 2.1.1: Presence of outliers: This examines if a data value in a series of values is extreme in relation to the other values in the series.

Table 1: Internal consistency outliers

Metric	Severity	Definition	
		National Level	Regional Level
Outliers (Analyze each indicator separately.)	**Extreme** (At least 3 standard deviations from the mean)	% of monthly regional unit values that are extreme outliers	# (%) of regional units in which ≥1 of the monthly regional unit values over the course of 1 year is an extreme outlier value
	Moderate (Between 2–3 standard deviations from the mean, or >3.5 on modified Z-score method)	% of regional unit values that are moderate outliers	# (%) of regional units in which ≥2 of the monthly regional unit values over the course of 1 year are moderate outliers

Outliers = Deviation from the mean

Table 2: Example of outliers in a given year for a certain indicator.

Woreda	Month												Total Outliers	% Outliers
	1	2	3	4	5	6	7	8	9	10	11	12		
A	2543	2482	2492	2574	3012	2709	3019	2750	3127	2841	2725	**2103**	1	8.30%
B	1184	1118	1195	1228	**1601**	1324	1322	**711**	1160	1178	1084	1112	2	16.70%
C	776	541	515	527	857	782	735	694	687	628	596	543	0	0%
D	3114	2931	2956	4637	**6288**	4340	3788	3939	3708	4035	3738	3606	1	8.30%
E	1382	1379	1134	1378	1417	1302	1415	1169	1369	1184	1207	1079	0	0%
National	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**4**	**6.70%**

The above table shows moderate outliers for a given indicator. There are four identified moderate outliers. They are highlighted in red. Three of the woredas have at least one occurrence of a monthly value that is a moderate outlier.

Nationally, this indicator is a percentage of values that are moderate outliers for the indicator. The numerator for the equation is the number of outliers across all administrative units [in this case, 4]. The denominator is the total number of expected reported values for the indicator for all the administrative units. That value is calculated by multiplying the total number of units (in the selected administrative unit level) with the expected number of reported values for one indicator for one administrative unit. In this case, we have 5 woredas and 12 expected monthly reported values per woreda for one indicator, so the denominator is 60 [5 × 12]. Thus, about 6.7% are moderate outliers [4/60 = 0.0666 × 100, or 6.7 %].

$$\text{Outlier for a certain indicator (\%)} = \frac{\text{\# of outliers across all administrative units}}{\text{\# Total number of expected report}}$$

Sub-nationally, see if you can calculate the number of outliers for each woreda. Count the woredas where there are two or more outliers (for moderate outliers) among the monthly values for the woreda [1]. Divide by the total number of administrative units [1/5 = 0.25 × 100 = 25%].

$$\text{Outlier for a certain indicator (\%)} = \frac{\text{\# of subnational unit with outliers}}{\text{\# Total number subnational units}}$$

Dimension 2.2.2: Consistency over time: The plausibility of reported results for selected programme indicators is examined in terms of the history of reporting of the indicators. Trends are evaluated to determine whether reported values are extreme in relation to other values reported during the year or over several years.

Table 3: Internal consistency: Trends over time

Metric	Definition	
	National Level	Regional Level
Trends/ Consistency over Time	Conduct one of the following, based on indicator's expected trend:	# (%) of woredas whose ratio of current year to predicted value (or current year to average of preceding 3 years) is at least ± 33% of national ratio.
(Analyze each indicator separately)	Compare current year to the value predicted from the trend in the 3 preceding years	
	Graphic depiction of trend to determine plausibility based on programmatic knowledge	

Table 4: Example of trends over time

Woreda	Year				Mean of Preceding 3 Years (2010-2012)	Ratio of 2013 to Mean of 2010-2012	% Difference between National and Woreda Ratios
	2010	2011	2012	2013			
A	30242	29543	26848	32377	28878	1.12	0.03
B	19343	17322	16232	18819	17632	1.07	0.08
C	7512	7701	7403	7881	7539	1.05	0.09
D	15355	15047	14788	25123	15063	1.67	**0.44**
E	25998	23965	24023	24259	24662	0.98	0.16
National	**98450**	**93578**	**89294**	**108459**	**93774**	**1.16**	

NB: Consistency trend: Comparison of woreda ratios to national ratios

Any di erence between woreda and national ratio that is 33% is highlighted in red.

Mean of preceding three years (2010, 2011, and 2012) is 93,774 [98,450 + 93,578 + 89,294)/3]

Ratio of current year to the mean of the past three years is 1.16 [108,459/93,774 ≈ 1.16].

The average ratio of 1.16 shows that there is an overall 16% increase in the service outputs for 2013 when compared to the average service outputs for the preceding three years of the indicator.

Regionally, try to evaluate each woreda, by calculating the ratio of the current year (2013) to the average of the previous three years (2010, 2011, and 2012). For example, the ratio for Woreda 1 is 1.12 [32,377/28,878].

Then calculate the % of difference between the national and woreda ratios for each woreda. For example, for woreda A:

= = 0.03 = 3.0%

The difference between the woreda ratio and the national ratio for Woreda A is less than 33%. However, there is a difference of approximately 44% for Woreda D between woreda ratio and the national ratio.

To calculate this indicator sub-nationally, all administrative units whose ratios are different from the country's ratio by ±33%, or more are counted. In this example, only Woreda D has a difference greater than ±33%. Therefore, 1 out of 5 woredas (20%) has a ratio that is more than 33% different from the national ratio.

Dimension 2.1.3: Consistency between indicators: Programme indicators which have a predictable relationship are examined to determine whether the expected relationship exists between those indicators. In other words, this process examines whether the observed relationship between the indicators, as depicted in the reported data, is that which is expected

Table 5: Internal Consistency: Comparing selected Related Indicators

Metric	Definition	
	National Level	**Regional Level**
Consistency among related indicators	**Maternal Health**: ANC1 – Syphilis test **(should not be negative)**	# (%) of regional units where there is an extreme difference (≥ ± 10%)
	Immunization: Penta3 dropout rate = (Penta1–Penta3)/Penta1 **(Should not be negative)**	# (%) of regional units with # of Penta3 immunizations >Penta1 immunizations (negative dropout)
	HIV/AIDS: (HIV positive pregnant women – HIV positive pregnant women who received ART) **(Should not be negative)**	# (%) of regional units where there is an extreme difference (≥ ± 10%)
	TB: (TB treatment success rate –TB Cure Rate) **(Should not be negative)**	# (%) of regional units where there is an extreme difference (≥ ± 10%)
	Malaria: # confirmed malaria cases reported - cases testing positive **(should be roughly equal)**	# (%) of regional units where there is an extreme difference (≥ ± 10%)

Table 6: Example: Internal Consistency

Region	ANC1	Syphilis test	Ratio of ANC1 to Syphilis test	% Difference between National &Regional Ratios
A	20995	18080	1.16	0.02
B	18923	16422	1.15	0.03
C	7682	6978	1.1	0.08
D	12663	9577	1.32	-0.14
E	18214	15491	1.18	0
National	**78477**	**66548**	**1.18**	

The annual number of pregnant women started on antenatal care each year (ANC1) should be roughly equal to the number of pregnant women who receive syphilis test in ANC, because all pregnant women should receive this test. First, we calculate the ratio of ANC1 to syphilis test for the national level, and then for each woreda. At the national level, the ratio of ANC1 to syphilis test is about 1.18 [78,477/66,548].

There is one woreda (D) with a ratio of ANC1 to syphilis test greater than 20%. We also see that the % difference between the national and woreda ratios for woreda D is more than 10%.

At the regional level, we can calculate the ratio of ANC1 to syphilis test and the % difference between the national and woreda ratios.

Dimension 2.1.4: Consistency of reported data and original records: This involves an assessment of the reporting accuracy for selected indicators through the review of source documents in health facilities. This element of internal consistency is measured by a data verification exercise which requires a record review to be conducted in a sample of health facilities. It is the only dimension of data quality that requires additional collection of primary data.

Dimension 2.2: External consistency with other data sources

The level of agreement between two sources of data measuring the same health indicator is assessed. The two sources of data usually compared are data flowing through the HMIS or the programme-specific information system and data from a periodic population-based survey. The HMIS can also be compared to pharmacy records or other types of data to ensure that the two sources fall within a similar range.

Table 7: External Consistency: Compare with Survey Results

Examples of Indicators	Definition	
	National Level	Regional Level
ANC 1st visit	Ratio of facility ANC1 coverage rates to survey ANC1 coverage rates	# (%) of aggregation units used for the most recent population-based survey, such as zone/state/region, whose ANC1 facility-based coverage rates and survey coverage rates differ by at least 33%
Penta3 vaccine	Ratio of Penta3 coverage rates from routine data to survey Penta3 coverage rates	# (%) of aggregation units used for the most recent population-based survey, such as zone/state/region, whose Penta3 facility-based coverage rates and survey coverage rates differ by at least 33%

Population-based surveys: Demographic and Health Survey (DHS), EPI Cluster survey.

- Indicator values are based on recall, referring to period before the survey (such as 5 years)
- Sampling error: confidence intervals

Table 8: Example: External Consistency

Woreda	Facility	Survey Coverage Rate	Ratio of Facility to Survey Rates	% Difference between Official and Alternate Denominator
A	1.05	0.95	1.10	10%
B	0.93	0.98	0.96	4%
C	1.39	0.90	1.54	54%
D	1.38	0.92	1.50	50%
E	0.76	0.95	0.80	20%
National	1.10	0.94	1.17	17%

NB: Comparison of HMIS and survey coverage rates for ANC1 Di erences 33% are highlighted in red.

If the HMIS is accurately detecting all ANC visits in the country (not just those limited to the public sector), and the denominators are accurate, the coverage rate for ANC1 derived from the HMIS should be very similar to the ANC1 coverage rate derived from population surveys. However, HMIS coverage rates are often different from survey coverage rates for the same indicator.

At the national level:

- The coverage rate from HMIS is 110%.
- The coverage rate from the most recent population-based survey is 94%.
- The ratio of the two coverage rates is: 1.17 [110%/94%].
- If the ratio is 1, it means that the two coverage rates are exactly the same.
- If the ratio is >1, it means that the HMIS coverage is higher than the survey coverage rate.
- If the ratio is <1, it means that the survey coverage rate is higher than the HMIS coverage rate.

The ratio of 1.17 shows that the two denominator values are fairly different, and there is about a 17% difference between the two values.

At the regional level, the ratio of denominators is calculated for each administrative unit. Woredas with at least 33% difference between their two denominators are flagged. Woredas C and D have more than 33% difference between their two ratios.

Dimension 2.3: External comparison of population data

The dimension on table examines two points:

- The adequacy of the population data used in the calculation of health indicators

The comparison of two different sources of population estimates (for which the values are calculated differently) to see the level of congruence between the two sources

Table 9: External Comparison of Population Data

Metric	Definition	
	National Level	Regional Level
Consistency of population projections	Ratio of population projection of live births from the Central Statistics Office to a United Nations live births projection for the country	NA
Consistency of denominator between program data & official government population statistics	Ratio of population projection for select indicator(s) from the census to values used by programs	# (%) of regional units where there is an extreme difference (e.g., ±10%) between the 2 denominators

Table 10: External Comparisons of Population Denominators

Woreda	Official Government Estimate for Live Births	Health Program Estimate for Live Births	Ratio of Official Government to Health Program Estimates
A	29855	29351	1.02
B	25023	30141	0.83
C	6893	7420	0.93
D	14556	14960	0.97
E	25233	25283	1
National	101560	107155	0.95

NB: Comparison of national and regional administrative unit ratios of o icial government live birth estimate
Administrative units with di erences ±10% are highlighted in red.

The above table shows the ratio of the number of live births from official government statistics nationally for the year of analysis to the value used by the selected health program.

Calculate the ratio of regional administrative unit 2014 live births to the value used by the selected health program; woreda B has a difference of 0.17 or 17%.

3. Completeness

All required data should be present and the medical/health record should contain all pertinent documents with complete and appropriate documentation.

Data Completeness on data recoding tools (Registers, cards/forms)

This refers all necessary data elements on registers/forms/cards should be filled immediately after provision of the service by the care provider.

- The cover page of integrated individual folder should contain all the necessary identifying data to uniquely identify an individual patient or client.

- For inpatients or clients received the service, the registers should contain all necessary information's accurately pertinent to the service provided and those include on registers.

- For all medical/health records, relevant forms are complete, with signatures and date of attendance.

Data completeness on reporting formats

- This refers the extent to which facility and woreda filled all data elements in the reports or data base for all reportable events. Health facilities are expected to fill a zero value in the reporting form even if the event doesn't happen in a defined reporting period.

$$\text{Completeness of data (\%)} = \frac{\text{\# values entered (not missing) in the report}}{\text{\# Total data elements in the report}}$$

$$\text{Completeness of reports (\%)} = \frac{\text{\# reports that are complete (all data elements filled out)}}{\text{\# Total reports available or received}}$$

The administrative unit will check all data element if they are left blank and take. Administrative health unit can calculate the proportion of data elements with zero value in monthly/quarterly service report from the total data element expected.

Explain!

N.B.: It is more important to calculate the content completeness for all reports separately to identify which report type has the gap and act on it accordingly. **Explain what this will tell us?**

Report Completeness

This helps to examine the total reports received from all health facilities from the total reports expected for a given period of time. All health posts and Health facilities are expected to send monthly (service and disease report), every quarter (Quarter service report) and annual service report once in a year. For HC and HSP with inpatient service IPD morbidity and mortality report is expected in monthly base

Report completeness (%) = # total reports available or received in a given period

Total reports expected with a given period

4. Timeliness

Information, especially clinical information, should be documented as an event occurs, treatment is performed or results noted. Delaying documentation could cause information to be omitted and errors recorded.

Example of timeliness

- A patient's identifying information is recorded at the time of first attendance and is readily available to identify the patient at any given time.
- On discharge or death of a patient in hospital, his or her medical records are processed and completed, coded and indexed within a specified time frame.
- All expected reports are ready within a specified time frame, having been checked, verified and sent to the next level with in a due date.

Report Timeliness (%) = #reports submitted or received on time

#total reports available or received

NB: All health facilities and administrative health units should have timeliness and completeness tracking logbook. If the facilities have electronic version of report tracking mechanism, they should use that one and keep the print-out as a record.

Sub Dimensions of data quality

5. Legibility

All data whether written, transcribed and/or printed should be readable.

Examples of legibility

- Handwritten demographic data are clearly written and readable.
- Handwritten notes on patient form, admission card, and any other medical records registers are clear, concise, readable and understandable.
- Handwritten National classification of diagnosis (NCoD) clear and easily understandable to transcribe in to Register

  In all medical/health records, cryptic codes or symbols cannot be used in either manual or electronic patient records.

If abbreviations are used, they are standard and understood by all health care professionals involved in the service being provided to the patient. Mostly this problem is seen at outpatient and inpatient department which are major source for clinical data and NCoD.

6. Accessibility

All necessary data are available when needed for patient care and for all other official purposes. The value of accurately recorded data is lost if it is not accessible.

Examples of accessibility

- Medical/health records are available when and where needed at all times.
- Abstracted data are available for review when and where needed.
- In an electronic patient record system, clinical information is readily available when needed.
- Statistical reports are accessible when required for Performance monitoring team, planning meetings and government requirements or for any official need.

7. Precision

This means that the data have sufficient detail. For example, an indicator requires the number of individuals who received HIV counseling and testing and received their test results by sex of the individual. An information system lacks precision if it is not designed to record the sex of the individual who received counseling and testing.

8. Confidentiality

Confidentiality means that clients are assured that their data will be maintained according to national and/or international standards for data. This means that personal data are not disclosed inappropriately, and that data in hard copy and electronic form are treated with appropriate levels of security (kept in locked cabinets and in password-protected files).

9. Integrity

Data have integrity when the systems used to generate them are protected from deliberate bias or manipulation for political or personal reasons.

10. Relevance

The data are logically connected with the matter in hand.

For instance, in using data to consider the program relevance, talking to others with knowledge of the program or target population who have in depth knowledge about the subject matter.

SECTION 4: DATA QUALITY ASSURANCE

Section duration: 2 days

Section Objectives

At the end of this section, participants will be able to:

- List the different types of data quality assurance techniques
- Understand and apply desk review of available data to check data quality
- Understand and apply the LQAS technique for checking reporting accuracy
- Explain and apply visual scanning as a tool to check for consistency of reports before/after conducting data entry
- Describe and apply RDQA as a self-assessment tool to monitor progress and evaluate the RHIS status.

Teaching Methods

- Lecture
- Group discussion
- Group presentation
- Exercise

Materials Needed

- PowerPoint presentations
- Projector
- Flip charts
- Markers

Activities

Question!

Recap data quality dimensions discussed in Section 2.

Activity: Discuss on quality assurance and data quality assurance.

- Write your responses on a flip chart.
- Actively participate on the brainstorming section.
- Compare your response to the displayed PPT

4.1. QUALITY ASSURANCE AND DATA QUALITY ASSURANCE

Quality Assurance: A program for the systematic monitoring and evaluation of the various aspects of a project, service, or facility (and taking actions accordingly) to ensure that standards of quality are being met" (Merriam-Webster Dictionary)

Data Quality Assurance: A systematic monitoring and evaluation of data to uncover inconsistencies in the data and data management system, and making necessary corrections to ensure quality of data

Data quality assessments help to improve data quality by uncovering hidden problems in data collection, aggregation, and transmission of priority indicator/data. Knowing about these problems allows health professionals and managers to develop data quality improvement plan.

There are different techniques used at facility and administrative levels to show the level of data quality and to take corrective measures.

Activity: Brainstorm on the types of data quality assurance tools.

- Write your responses on a flip chart.
- Actively participate on the brainstorming session.

4.2. TECHNIQUES OF DATA QUALITY ASSURANCE

The following methodology shall be applied to assure data quality at service delivery and intermediate health administration units

- A desk review of the data that have been reported to national level whereby the quality of aggregate reported data for recommended program indicators is examined using standardized data quality metrics;
- Health facility assessment

- Data Quality Checks using LQAS method
- Other health facility assessments to conduct data verification and an evaluation of the adequacy of the information system to produce quality data (system assessment).

x Administrative health unit level data quality assessment

- Routine Data Quality Assessment (RDQA)
- Data Quality Audit (DQA)
- Performance of Routine Information System Management (PRISM

Data Quality assurance techniques:

- Data Quality Desk review
- Data Quality Checks using Lots quality assurance sampling (LQAS),
- Routine data quality Audit (RDQA),
- Data quality Audit (DQA)
- Performance of routine information system management (PRISM),
- Visual scanning (Quantitative and qualitative data check) are some of the tools used to assess the performance of HIS.

Figure 2 Major Differences among LQAS, DQA, RDQA and PRISM

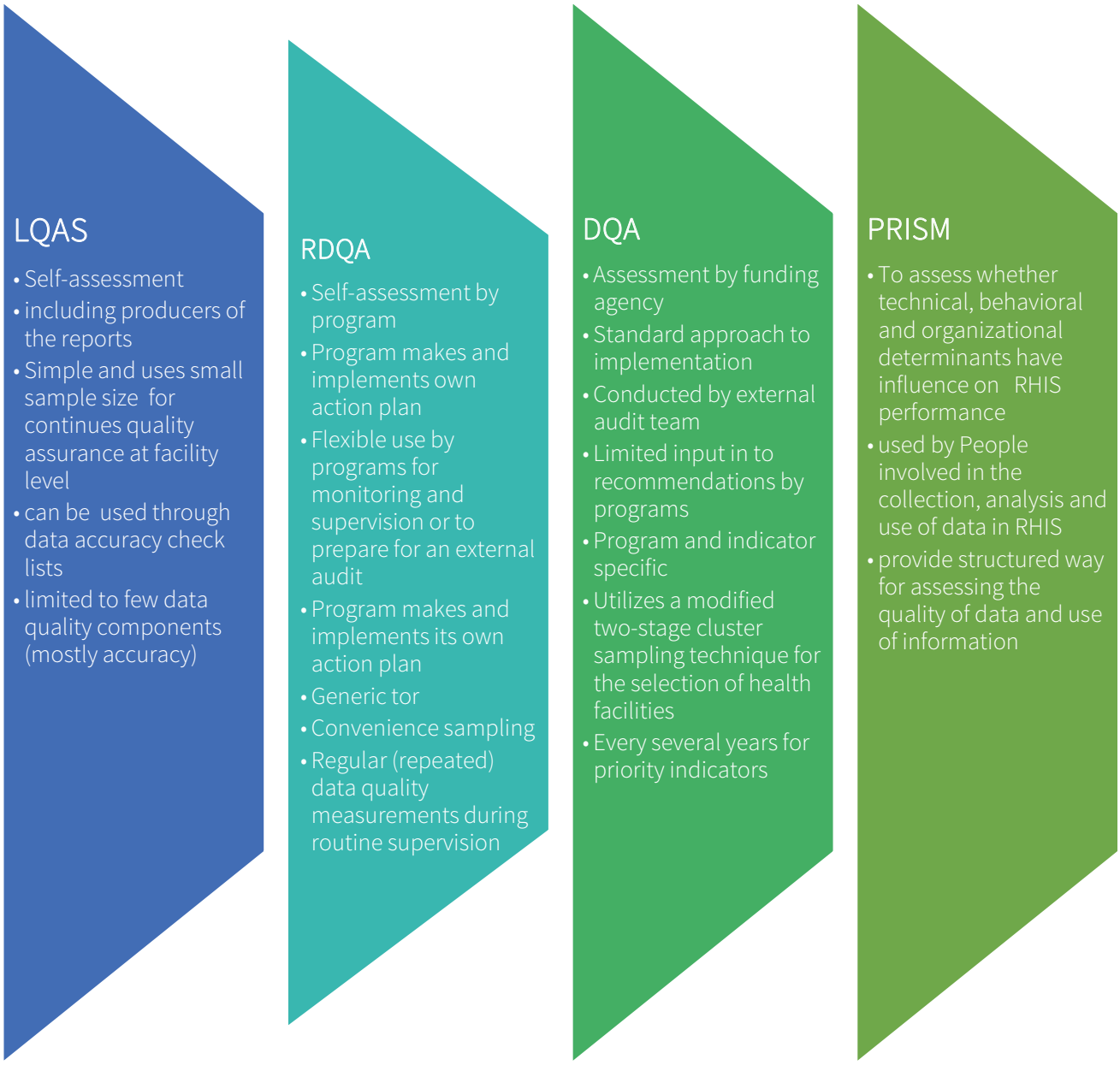
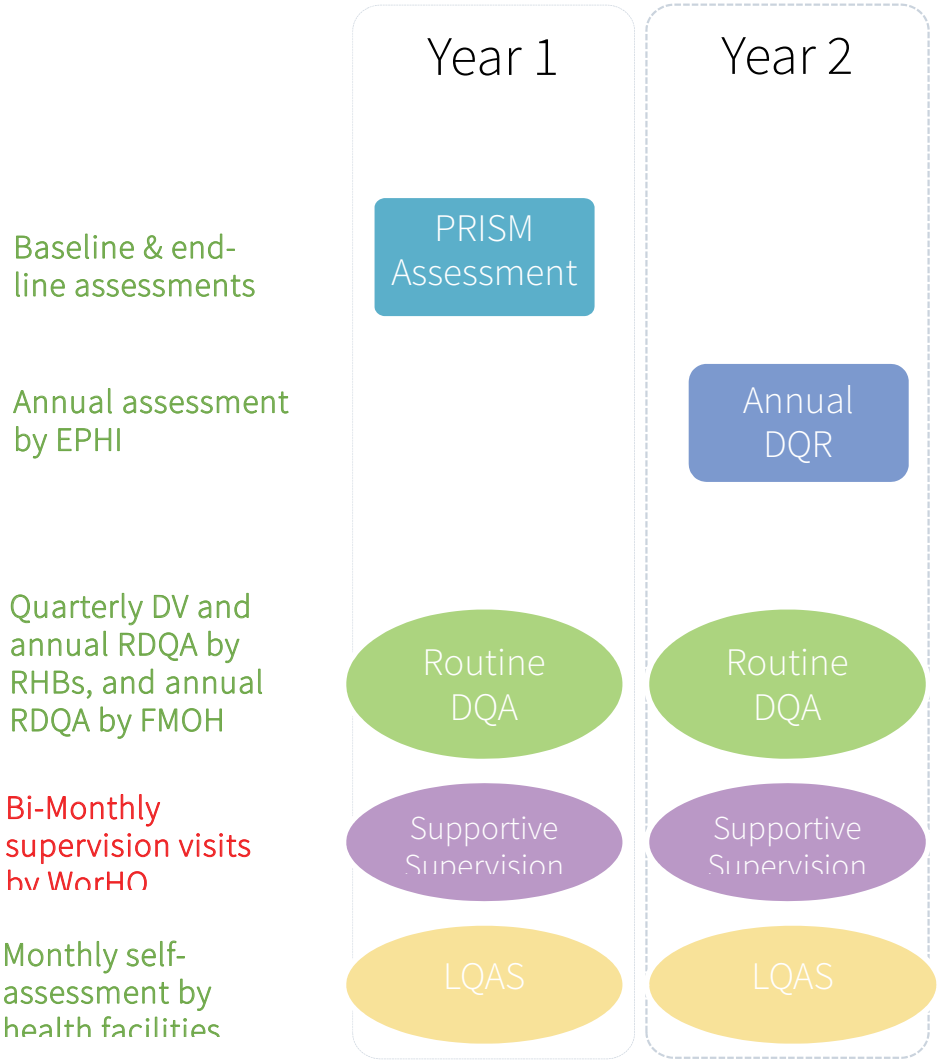


Figure 3 Ethiopian Data Quality Assurance Timeline



4.2.1 Data Quality Desk review

Description

The desk review examines a core set of tracer indicators selected across program areas in relation to these dimensions. The desk review requires monthly or quarterly data by subnational administrative area for the most recent reporting year and annual aggregated data for the selected indicators for the last three reporting years.

This cross-cutting analysis of the recommended program indicators across quality dimensions quantifies problems of data quality according to individual program areas but also provides valuable information on the overall adequacy of health-facility data to support planning and annual monitoring.

The desk review compares the performance of the country information system with recommended benchmarks for quality, and flags for further review any subnational administrative units which fail to attain the benchmark. User-defined benchmarks can be established at the discretion of assessment planners.

The desk review has two levels of data quality assessment:

- an assessment of each indicator aggregated to the national level;
- The performance of subnational units (e.g. districts or Zones/regions) for the selected indicators.

Who

FMOH and RHBs /ZHDs. This desk review is expected to be done at the M&E units and feedback on the findings should be communicated back for further.

Frequency

WHO recommends that the data quality desk review to be conducted annually. As many of the consistency metrics require annual data, the FMOH also recommends conducting this review annually.

Data quality dimensions addressed

Dimension 2: Consistency

Dimension 3.1: Internal consistency of reported data; (except Consistency of reported data and original records):

Dimension 3.2: external consistency

Dimension 3.3: external comparisons of population data

Dimension 3: Completeness (except Data Completeness on data recoding tools-Registers, cards/ forms)

Dimension 4: Timeliness

Data requirement

The desk review requires monthly or quarterly data by subnational administrative area for the most recent reporting year and annual aggregated data for the selected indicators for the last three reporting years.

Information on submitted aggregate reports and when they were received will be required in order to evaluate completeness and timeliness of reporting.

Other data requirements include denominator data for calculating coverage rates for the selected indicators and survey results (and their standard errors) from the most recent population-based survey – such as the Demographic and Health Surveys (DHS) and immunization coverage surveys.

How

Doing data quality desk review manually is very cumbersome as well as challenging and it also needs advanced data analysis skills. The Ethiopian MOH has customized DHIS 2 to include dashboards for analyzing and displaying the above stated data quality metrics. Detailed discussion and hands on training will be provided under Section 4.

4.2.2. Lot Quality Assurance Sampling (LQAS)

Lot Quality Assurance Sampling (LQAS) - is a technique useful for assessing whether the desired level of data accuracy has been achieved by comparing data in relevant record forms (i.e. registers or tallies) and the HMIS reports.

Description:

It is a technique useful for assessing whether the desired level of reporting accuracy has been achieved by comparing data in relevant record forms (i.e. registers or tallies) and HMIS reports. The data that is compiled in databases and reporting forms is accurate and reflect no inconsistency between what is in registers and what is in databases/reporting forms at facility level. Similarly, when data entered in the computers, there is no inconsistency between reporting forms and computer file.

The LQAS method will be used to check reporting accuracy at Health Facility level. The Health facilities will maintain a registry to record the data consistency check results and to look the trend of the data quality improvement.

This is a method for testing hypothesis related with the level of HMIS data quality whether it is achieved or not. It uses a sample size of 12 data elements and tries to check the reporting accuracy.

If the number of sampled data elements not meeting the standard exceeds a pre-determined criterion (decision rule), then the lot is rejected or considered not achieving the desired level of pre-set standard. Decision rule table is used for determining whether the pre-set criterion is met or not. Comparison of LQAS results over time can indicate the level of change.

Who

Health facilities (Hospital, health center and health posts).

Frequency

Monthly

Data quality dimension addressed

Dimension 3.1: Internal consistency of reported data; (Consistency of reported data and original records)

How

Steps to carryout LQAS

Step 1 Decide the month for which you want to do the data accuracy check.

Step 2 Pre-fix the level of data accuracy that you are expecting, e.g. 85% or 90% etc.

Step 3 Put serial numbers against the data elements (not disaggregation) in the Service Delivery or Disease Report that you want to include in the data accuracy check

Step 4 Generate twelve random numbers using Excel program. These random numbers represent the serial numbers of the data elements included in the data accuracy check. Note them in Column of the Data Accuracy Check Sheet. This is to ensure representation of all data elements by giving equal chance to all data elements.

Step 5 List down the selected data elements from the report on to the Data Accuracy Check Sheet in Column 2 and Column 3

Step 6 Write down the reported figures from the Monthly HMIS Report for the selected data elements in the Column 4 of the Data Accuracy Check Sheet.

Note: In case of Health Post, figures for the selected data elements from the Tally Sheet will be compared with recounted figures from the Family Folders. Therefore, record the figures for the selected data elements from the Tally Sheet in Column 5

Step 7 Recount the figure from the corresponding registers and note the figures on Column 5 of the LQAS check-sheet

Step 8 If the figures for a particular data element match or do not match put “yes” or “no” accordingly in Column 6 or Column 7 respectively.

Step 9 Count the total number of “yes” and “no” at the end of the table

Step 10 Match the total number of “yes” with the LQAS Decision Rule table and determine the level of data accuracy achieving the expected target or not.

Please complete the steps on the Handout.

Questions

- In your view, what should be the desired HMIS data accuracy level?
- In order for the HMIS report to meet the desired accuracy level, how many data elements would completely match? (Ask them to find the desired number of matches in the "Decision Rule" table)
- How many data elements on the handout show that they match?
- What is the data accuracy level achieved?
- Does that level meet the desired data accuracy level?
- Invite questions from the participants and clarify accordingly.

Handout: LQAS Data Accuracy Check sheet

Random No.	Reporting Element	Figures from			Does figure from source documents match?	
		Report	Tally	Register	Yes	No
(1)	(3)	(4)	(5)	(6)	(7)	(8)
1	Repeat Acceptors	14		14	X	
2	Deliveries attended by skilled health personnel	52		32		X
10	Fully Immunized infants <1 yrs. of age	12	15	15		X
18	2-5 yrs. age group who de-wormed	26		26	X	
8	Measles doses given <1years of age	8	8	8	X	
20	Live birth	32		28		X
5	Number of newborns weighed	28		28	X	
35	Number of weights recorded with severe malnutrition	78	80	80		X
40	Pregnant mothers linked based on option B+ for the first time	0		0	X	
65	Early PNC within 0-48 hours	4		4	X	
5	Vitamin-A supplementation for 6-59 months of age	2		2	X	
12	Early neonatal death in the first 24hr	11		14	X	
Total Yes or No					**8**	**4**

Decision Rules

Decision Rules for sample Sizes of 12 and Coverage Targets /Average of 20-95%															
Sample size	Average Coverage (baselines)/Annual Coverage Targets (monitoring and Evaluations)														
	<20%	20%	25%	30%	35%	40%	45%	55%	60%	65%	70%	75%	80%	85%	90%	95%
12	N/A	1	1	2	2	3	4	5	6	7	7	8	8	9	10	11

The HMIS focal should do LQAS check by repeating the same procedure after having the revised report. However, the first LQAS score should be reported in the monthly report format and the health facility should keep the record of both LQAS accuracy sheet on PMT minute book (Data quality Log book). The Health facilities should monitor the trend of LQAS across months to see the changes overtime.

Please note that Health Facilities will maintain a registry to record the data accuracy check results. The HMIS focal persons will also use it for recording the data accuracy check during their supportive supervision visits.

Question?

What actions would be necessary if they find that the data accuracy at a health facility is not of the desired level?

Activity: Checking Data Quality

Step 1: Conduct data quality check at facility level. We are checking how many mistakes are made during the transfer of data from registers to monthly reporting forms. Thus, you need various registers, a monthly reporting form.

For this exercise, please use;

- Copies of outpatient, under-5, antenatal, postnatal, and family planning registers;
- Monthly reporting form

Step 2: Checking Data Quality

- Select randomly any 12 data points—with numbers-- from the monthly report form. Enter them into the first column of the data quality check.
- Copy the number from the monthly report form into the second column of the data quality checklist under the heading of monthly report.
- Calculate the total number of selected data items and enter that number into the third column of the data quality checklist, under the heading register.

- If the numbers are same in columns 2 and 3, enter “yes” in column 4, otherwise “no.”
- Calculate total matched and mismatched numbers and write under row of total. Total matched numbers are the accurate number.

4.2.3. Visual Scanning (Eye Balling)

It is a simple method used at health facility to check for consistency of reports before/after conducting data entry. The PMT members sit together and look across each line and then from top to bottom to identify **missing data values**, **unexpected fluctuations beyond maximum/minimum values**, **inconsistencies between linked data elements**, and for **mathematical errors.**

Examples:

- Family planning acceptors by age and method disaggregation
- Antenatal first attendance by gestational and age disaggregation
- Delivery attended by skilled health personnel vs Sum of still birth and live birth

Frequency:

Whenever report is generated

Data quality dimensions addressed:

- Presence of outlier
- Data completeness
- Internal consistency between indicators

4.2.4. Routine Data Quality Assessment (RDQA)

Routine Data Quality Assessment (RDQA) tool helps to:

- Perform data accuracy at administrative level by enabling quantitative comparison of recounted data to reported data
- Assess if intermediate aggregation sites are collecting and reporting data accurately by providing a “Verification Factor” i.e. level of under or over reporting, if any, for the HMIS data items studied.

RDQA is an assessment technique that can be used to self-assess and to monitor progress and evaluate the RHIS status. Unlike to LQAS, the RDQA help the Health facilities and administrative health units to verify reported data against to source documents and to look RHIS system implementation. It is a simpler version of the DQA. Each level of the data management system has a role to play and specific responsibilities in ensuring data quality throughout the system. The RDQA tool should be applied regularly to monitor the trend in data quality. It is recommended to be implemented quarterly by administrative health unit and Health facilities can use for self-assessment purpose in a much-customized way.

Objective of RDQA:

By using the RDQA tool, we can achieve three main objectives.

1. **Verify rapidly**
    - the quality of reported data for key indicators at selected sites;
    - the ability of data management systems to collect, manage, and report good-quality data
2. **Implement**
    - corrective measures with action plans for strengthening the data management and reporting system
    - improving data quality
3. **Monitor**
    - capacity improvements and performance of the data management and reporting system to produce good-quality data

Activity: Discuss in groups about the importance of RDQA

Importance and Components of RDQA

Importance of RDQA

1. Routine data quality checks as part of on-going supervision
    - Routine data quality checks can be included in already planned supervision visits at the service delivery sites.
2. Initial and follow-up assessments of data management and reporting systems
    - Repeated assessments (e.g., biannually or annually) of a system's ability to collect and report quality data at all levels can be used to identify gaps and monitor necessary improvements.
3. Strengthening program staff's capacity in data management and reporting
    - M&E staff can be trained on the RDQA and be sensitized to the need to strengthen the key functional areas linked to data management and reporting in order to produce quality data
4. Preparation for a formal data quality audit

- The RDQA tool can help identify data quality issues and areas of weakness in the data management and reporting system that would need to be strengthened to increase readiness for a formal data quality audit

5. External assessment by partners of the quality of data

   - Such use of the RDQA for external assessments could be more frequent, more streamlined and less resource intensive than comprehensive data quality audits that use the DQA version for auditing.

Components of RDQA

RDQA tool has two key components, which are data verification and system assessment.

1. **Data Verification part:** facilitates a quantitative comparison of recounted to reported data and a review of the timeliness, completeness and availability of reports.

The purpose of this part of the RDQA is to assess if:

- Service delivery and intermediate aggregation sites are collecting and reporting data accurately, completely, and on time, and
- Whether the data agrees with reported results from other data sources.

2. **System assessment:** this part enables qualitative assessment of the relative strengths &weaknesses of functional areas of a data management and reporting system. The purpose of assessing the data management and reporting system is to identify potential threats to data quality posed by the design and implementation of data management and reporting systems.

Basic implementation and assessment areas of RDQA

RDQA tool can be implemented at any or all levels of the data management and reporting system, M&E Unit; intermediate aggregation levels (e.g. region and woreda); and/or service delivery points.

RDQA tool has six parts that help to asses and improve the RHIS performance, which are data verification, system assessment, interpretation of outputs, development of action plans, dissemination of results, and on-going monitoring.

RDQA focus in two major assessment methods: 1) **Documentation review** -describe answering yes/no questions to whether the source documents required for the assessment are available, completed and within the required reporting period. 2) **Data Verification** -helps to check whether the indicator of interest found in the periodic summary report against an alternative data source. The degree to which the two sources match is an indication of good data quality.

Steps followed to conduct RDQA

Step 1: Determine the Purpose of RDQA

Discuss in small group about the purpose of RDQA

Remind the participants on some of RDQA purpose

- Routine data quality checks as part of on-going supervision
- Initial and follow-up assessments of data management and reporting systems
- Strengthening program staff's capacity in data management and reporting
- Preparation for a formal data quality audit
- External assessment by partners of the quality of data

Step 2: Selection of study sites

Once the purpose of RDQA has been determined, the second step in the RDQA is to decide what levels of the data management and reporting system will be included in the assessment (service delivery sites, intermediate aggregation levels (e.g. regions, woredas), and/or the central M&E unit.). It is not necessary to visit all the reporting sites in a given Program to determine the quality of the data or how HIS system function. Random sampling techniques can be used to select a representative group of sites whose data quality is indicative of data quality for the whole program.

A. Types of sampling methods for selecting sites for the RDQA

There are different sampling methods for selecting sites for the RDQA. The sampling methods include purposive, restricted site design, stratified random, random, and cluster sampling methods. Most recommended method is two stage random sampling and the objective of the study is the base for selecting the type of sampling. Please refer Annex-x for more detailed instructions about sampling.

B. Determining the number of sites at National M&E unit

Study sites are widely distributed and the various administrative levels are not of equal size, hence the need to have a sampling frame that involves selection of clusters accordingly. All regions will be involved in the RDQA and the primary sampling unit for the sampling is cluster or woredas which refer to the administrative or political or geographic unit in which Service delivery sites are located. A probability proportionate to size (PPS) will be used to derive the total set of clusters (woredas) from each region that the assessment will include. Then the actual Clusters (woredas) are selected in the first stage using systematic random sampling, where clusters having active HMIS reporting system are listed in a sampling frame by region. In the second stage, Service delivery Sites from selected clusters are chosen using stratified random sampling where the service delivery sites are stratified on volume of service (or OPD attendance per capita (<=0.5 and >0.5). And because of financial and logistic feasibility, two health centers from each stratum and one hospital will be selected randomly from each selected woreda.

2. Determine the number of clusters and sites: To estimate the sample size of the clusters (woredas) from the regions a single population proportion formula should be used:

$$n = \frac{p\,(1-p)\,z1-\alpha/2}{S^2}$$

Where:

p= the estimated proportion of data quality (If a previous study exists, p will be the accuracy level of the indicator which provide the highest sample size or p will be 50% if no study exists)

z1-α/2 = the z score corresponding to the probability with which it is desirable to be able to conclude that an observed change of size could not have occurred by chance (α= 0.05 (z1-α/2= 1.96) and from the precision or margin of error denoted by (s) found that 0.05.

If N (the total number of clusters or woredas) < 10,000, a correction formula will be used.

nf= n/1+ (n/N)

C. Determining the number of sites at Regional level:

The above stated sampling methodologies can be employed to select the appropriate number of sites and clusters based on the objectives of the assessment. Precise estimates of data quality require a large number of clusters and sites. Often it isn't necessary to have a statistically robust estimate of accuracy. That is, it is sufficient to have a reasonable estimate of the accuracy of reporting to direct system strengthening measures and build capacity. A reasonable estimate requires far fewer sites and is more practical in terms

of resources. Generally, 12 sites sampled from within 4 clusters (3 sites each) are sufficient to gain an understanding of the quality of the data and the corrective measures required.

The Ethiopian MOH recommends the following sample size and methodology for RDQA (especially for DV):

1. **In regions with zones:**
    - Randomly select 4 zones
    - From each of the selected zones, randomly select three Woredas
    - From selected Woredas, select randomly one health center or hospital
2. **In Regions without zones**
    - Randomly select 4 Woredas
    - From each selected Woredas, randomly select three health centers or hospitals
3. **For Zonal level**
    - Randomly select 4 Woredas
    - From selected each Woredas, select randomly three health centers or hospitals
4. **For Woreda level**
    - Use census of all health centers and hospitals in the Woreda

D. Frequency

It is suggested that frequency of RDQA has to be based on the objective of the assessment and the level of the organization conducting it. Accordingly, the data verification part has to be done quarterly integrating it with supportive supervision visits by organizations at all levels; whereas it is recommended that a comprehensive RDQA (Data verification and system assessment) should be done annually by Federal or regional level coordinating bodies. It is also important to clearly identify the reporting period associated with the indicator(s) to be assessed. Ideally, the time period should correspond to the most recent relevant reporting period or schedule in HMIS.

Level	Data Verification	Full RDQA
FMOH	Bi-annually	Annually
RHB	Quarterly	Annually
WoHO	Every two months	NA
Health Facilities	NA	NA

Step 3: Selection of Indicators and data source

Determination of indicators and reporting period that should be included in the assessment is also an important step in RDQA. It is recommended that up to two indicators be selected within a Disease/Health Area and that, if multiple Diseases/Health Areas are included in a Data Quality assessment, that a maximum of four indicators can be included. More than four indicators could lead to an excessive number of sites to be evaluated.

The criteria for selecting the indicators for the RDQA could be the following:

1. Must review indicators: Indicators that should be selected first depending on the indicator's national and global importance/ priority.
2. Relative magnitude of the indicators: The amount of budget and activity associated with the indicator(s).
3. Case by Case Purposive Selection: Indicators for which data quality questions exist and the government wants to be routinely verified. Those reasons should be documented as justification for inclusion.

Step 4: Conduct Site Visits

Selected sites should be notified prior to the visit for the data quality assessment. This notification is important in order for appropriate staff to be available to answer the questions in the checklist and to facilitate the data verification by providing access to relevant source documents.

The team should be seated with facility in-charge and other management members and explain the objective of the assessment before starting the formal data collection. The data collecting team may spend half day at one health facility just by filling the checklist. During the site visits, the relevant sections of the appropriate checklists in the Excel file are filled out (e.g. the service site checklist at service sites, etc.). These checklists are completed following interviews of relevant staff and reviews of site documentation. The copy should be given for the facility to look their gaps and take corrective measures even before the release of official report.

The tool and its components

RDQA tool has 19 worksheets and the first two sections gives general information on how to use the tool and the rest help in data collection and data analysis.

Tell participants that this training will focus on Data Verification component and a separate full RDQA train
will be provided for those who will be involved in the comprehensive assessment.

1) Data Verification

The purpose is to assess, on a limited scale, if service delivery and intermediate aggregation sites are collecting and reporting data to measure the indicator(s) accurately and on time — and to cross-check the reported results with other data sources. To do this, the RDQA will determine if a sample of Service Delivery Sites have accurately recorded the activity related to the selected indicator(s) on source documents.

The data verification exercise will take place in two stages:

1. In-depth verifications at the Service Delivery Sites;

1.1 Verify reported data against recounted from registers

Example

Indicators	Description	HF1	HF 2	HF3	HF4	HF5	HF6	HF7	ΣA / ΣB	VF= A/B
ANC4	Recounted=A	10	50	70	20	30	40	20	240	0.89
	Reported=B	12	65	70	20	25	45	30	267	
SBA	Recounted=A	111	44	2	20	10	9	15	211	0.93
	Reported=B	121	43	0	12	25	9	15	225	
Penta 3	Recounted=A	25	45	30	12	20	10	0	142	0.83
	Reported=B	38	59	30	16	15	13	0	171	
Currently on ART	Recounted=A	10	22	10	5	40	19	20	126	1.94
	Reported=B	0	12	4	5	32	12	0	65	
Meseals	Recounted=A	20	55	34	14	45	25	27	220	0.79
	Reported=B	12	42	23	22	95	36	47	277	
TB all forms	Recounted=A	41	71	29	78	9	1	12	241	1.14
	Reported=B	29	36	34	80	6	10	17	212	

1.2 **Verify the primary source of data (Medical records) against the secondary source of data (registers):** The purpose of this verification process is to measure the level of under reporting by comparing data elements from medical records and registers. It is a method of randomly selecting 10-20 medical records from the Card room and verifying if all the data elements that are supposed to be recorded are captured in the register. We can summarize the data for selected medical records as complete or incomplete based on the number of data elements recorded for the latest visit that matched between the medical record and the register.

Instructions

1. Randomly select 10 sample medical record numbers from the central register from the list of patients who were seen in the last three days
2. Write the medical record number of each card in the first column
3. For each card, identify the latest visit date
4. Identify the register(s) based on the diagnosis (N.B check if the service delivery unit and is written on the summary sheet)
5. Match if all the relevant data elements in the card are recorded in the register
    - x If all the data elements in the medical record are recorded in the register, mark that card as "Complete" in the second column
    - x If not, Mark as "Incomplete"
6. Count the number of complete medical records and divide it with the total sampled medical records.
7. Analyze the level of under-reporting based on the decision table below

<50%	50-75%	75-85%	>85%
Catastrophic level of under-reporting	Severe under-reporting	Moderate level of under-reporting	Acceptable

Under-reporting Computing Sheet

Medical record #	**Complete** (all the data elements in the medical record are recorded in the register)	**Incomplete** (one or more data elements in the medical record were not recorded in the register)
00057	X	
00119		X
00362	x	
00007	x	
00137	x	
00999		X
01120		X
01070	x	
00082	x	
02200	x	
Total	7	3

1.3 Cross-check secondary data source (Registers) with the primary data source (Medical records).

The purpose of this verification process is to measure the consistency of register and the medical record. It particularly measures the level of over reporting.

Instructions

1. Select two core data elements from the sample indicators selected for data verification
2. From the respective register, select 5-10% of the total recoded data within the reporting period

   Example if total SBA recoded in the register is 200 will take 5% which is 10 to verify the data at me record room.

3. Take out the medical records for the sampled cases. To randomly select medical records, divide the total number recorded by the required number of the sample (e.g. 10) to obtain the sampling interval.

   In this Example the sample interval will be 20 i.e. we will take every 20th patients.

4. Match the recorded data in the register against the medical record
   a. If the recorded data element in the register is found in the medical record, mark that card as "Matched"
   b. If not, Mark as "Not Matched". This also include if the medical record is not physically available in the card room, it is also considered as not matched.
5. For each data element, analyze the level of consistency based on the decision tree below

<50%	50-75%	75-85%	>85%
Catastrophic level of inconsistency	Severe inconsistency	Moderate level of inconsistency	Acceptable

Consistency Computing Sheet

Data Element for selected indicators	Medical record #	**Matched** (recorded data element in the register is found in the medical record)	**Not Matched** (data element recorded in the register is not found in the medical record orthe medical record is not physically available)
ANC 4	00057	X	
	00119		x
	00362	X	
	00007	X	
	00137	X	
	Total	4	1
SBA	00999		X
	01120		X
	01070	X	
	00082	X	
	02200	X	
	Total	3	2

Possible reasons for inconsistency between the register and the medical record.

1. Over reporting
2. Data falsification
3. Loss of medical record
4. Service provision without medical record

1.4 Community Level data verification

- From the matched core data elements of the selected priority indicators during cross-checking of secondary data source (Registers) with the primary data source (Medical records), randomly select 50% of the medical records or a minimum of five (whichever is bigger) and verify whether the patients or clients have accessed the service within the specified period.
- The verification should be done via telephone or house to house visit. The house visit should be accompanied by HEWs for easy access to the house of the clients
- The team should document basic demographic information (Name, Kebele, got House number, phone number), date of the service provided, and type of service provided before departure to household level verification.

Key Points for Community Verification

- Objectives of the community verification should be explicitly explained before the process started
- Make sure that the selected indicators for community verification are not sensitive.
- Verification at community level should not be done by proxy. The actual client should be contacted.

Follow-up verifications at the Intermediate Aggregation Levels and at the program/ project M&E Unit. (Will be discussed on intermediate levels RDQA form)

2. Data verification at intermediary aggregation level

It will help to see if data has been correctly aggregated and/or otherwise manipulated as it is submitted from the initial Service Delivery Sites through intermediary levels to the program/project M&E Unit.

It has two sections:

a) **Recounting reported data from service delivery units (Health facilities)**

- Recount results from the periodic reports sent from service sites (Health facilities) to the Woreda and compare to the value reported by the Woreda. (This is more applicable if the report is submitted manually. If it is electronic there will not be room for data manipulation at intermediate level, hence no data verification is needed)

b) **Reporting performance**

- Review availability, completeness, and timeliness of reports from all Service Delivery Sites. How many reports should there have been from all Sites? How many are there? Were they received on time? Are they complete?

Case Study: Data Verification and Reporting Performance

As part of the RDQA Assessment in Ethiopia, the FMoH would like to verify the data accuracy and reporting performance of the Family Planning program. The indicator selected was "Contraceptive Acceptance Rate."

The Woredas and health facilities that were selected to be included in the RDQA assessment were assigned across several assessment teams. Team #5 was responsible for conducting the assessment at Endegagn Woreda Health O ice in Gurage Zone.

Endegagn Woreda Health O ice is expected to receive reports from 3 health facilities (1 primary hospital and 2 health centers) on a monthly basis. The reports should arrive by the twenty-sixth day of the month. The reporting period selected for verification is December 2017.

Using the reports received (see below), verify the data and calculate the reporting performance at the woreda level for the indicator "Contraceptive Acceptance Rate." Please note that recounted figures for the same period for Jane HC, Dinkula HC, and Dinkula hospital are 38, 62 and 80 respectively.

Specifically, calculate the following data quality indicators:

- Accuracy (explain there is any over or under reporting)
- Reporting completeness (availability of reports)
- Data completeness (reports with data elements filledout)
- Timeliness
- Internal data consistency

Federal Minstry of Health Hospital/Health Center/Clinic/Center monthly Service Delivery Report

Activity	Number
Reproductive and Maternal Health Contraceptive acceptance rate	
Total new and repeat acceptors, disaggregated by age	40
New acceptors by age	13
10-14 yr	0
15-19 yr	4
20 - 24 yr	7
25 -29 yr	2
30 -49 yr	0
Repeat acceptors by age	28
10-14 yr	2
15-19 yr	5
20 - 24 yr	12
25 -29 yr	6
30 -49 yr	5
Total new and repeat acceptors, disaggregated by method	52
New acceptors, by method	19
Oral contraceptives	9
Injectable	7
Implants	2
IUCD	0
Vasectomy	0
Tubal ligation	0
Others	0
Repeat acceptors, by method	33
Oral contraceptives	13
Injectable	8
Implants	9
IUCD	3
Vasectomy	0
Tubal ligation	0
Others	0

Asena Bireda
Name

Tir 01, 2010
Date Submitted

Federal Minstry of Health Hospital/Health Center/Clinic/Center monthly Service Delivery Report	
Region **SNNPR** Zone **Gurage** Woreda **Endegagn** Name of the health facility **Dinkula HSP**	
year **2010** month **Tahisas**	
Activity	**Number**
Reproductive and Maternal Health	
Contraceptive acceptance rate	
Total new and repeat acceptors, disaggregated by age	82
New acceptors by age	35
10-14 yr	0
15-19 yr	12
20 - 24 yr	7
25 -29 yr	10
30 -49 yr	6
Repeat acceptors by age	47
10-14 yr	0
15-19 yr	7
20 - 24 yr	15
25 -29 yr	23
30 -49 yr	2
Total new and repeat acceptors, disaggregated by method	81
New acceptors, by method	35
Oral contraceptives	16
Injectable	6
Implants	9
IUCD	3
Vasectomy	0
Tubal ligation	0
Others	0
Repeat acceptors, by method	47
Oral contraceptives	7
Injectable	19
Implants	10
IUCD	11
Vasectomy	0
Tubal ligation	0
Others	0
Abebech Tofiq	Tahisas 26, 2010
Name	**Date Submitted**

Federal Minstry of Health Hospital/Health Center/Clinic/Center monthly Service Delivery Report	
Region **SNNPR** Zone **Gurage** Woreda **Endegagn** Name of the health facility **Dinkula HC**	
year **2010** month **Tahisas**	
Activity	**Number**
Reproductive and Maternal Health	
Contraceptive acceptance rate	
Total new and repeat acceptors, disaggregated by age	35
New acceptors by age	16
10-14 yr	0
15-19 yr	4
20 - 24 yr	7
25 -29 yr	3
30 -49 yr	2
Repeat acceptors by age	47
10-14 yr	0
15-19 yr	7
20 - 24 yr	15
25 -29 yr	23
30 -49 yr	2
Total new and repeat acceptors, disaggregated by method	
New acceptors, by method	16
Oral contraceptives	6
Injectable	9
Implants	1
IUCD	0
Vasectomy	0
Tubal ligation	0
Others	0
Repeat acceptors, by method	47
Oral contraceptives	7
Injectable	19
Implants	10
IUCD	11
Vasectomy	0
Tubal ligation	0
Others	0
Shemsu Dejene	Tahisas 28, 2010
Name	**Date Submitted**

2) Systems Assessment

The purpose of the system assessment is to identify potential challenges to data quality created by the data management and reporting systems at:

1. the service delivery sites, and
2. Any intermediary aggregation level (at which reports from service delivery Sites are aggregated prior to being sent to the M&E Unit).

The system assessment has six areas to be checked at service delivery sites and intermediate aggregation level:

1.

Summary Tables

To simplify the process of reviewing feedback from various sites or at various levels, the latest version of the RDQA tool has been updated to include worksheets with tables that automatically populate with the comments and remarks about the responses to the RDQA questions. The RDQA workbooks summarize results for Data verification quantitative comments, System assessment comments and detail of system assessment.

Step 6: Develop a system strengthening plan, including follow-up actions.

Based on the findings at each site the team will develop specific action plan at level and provide feedback. In addition to this after reviewing the overall results the RDQA team should create **action plans** to improve data quality and system assessment based on the objective of the study.

Engaging the team members, will create ownership of the plan and get the direct insights from the people on the field. Decisions on where to invest resources for system strengthening should be based on the relative strengths and weakness of the different functional areas of the reporting system identified via the RDQA, as well as consideration of practicality and feasibility.

Table x: Frequency of data quality techniques applied by administrative unit level and health facility

	DESK REVIEW	RDQA		LQAS	Eyeballing
		DV	Complete		
FMOH	Annually	Bi-annually	Annually	Not Applicable	Monthly
RHB	Annually	Bi-annually	Annually	Not Applicable	Monthly
WoRHO	Annually	Quarterly	Not Applicable	Not Applicable	Monthly
Health facilities	Not Applicable	Not Applicable	Not Applicable	Monthly	Monthly

Table XX: Data quality techniques applied and data quality dimension addressed by administrative unit level and health facility DQREPHI

SECTION 5: USING DHIS2 TO IMPROVE DATA QUALITY

Duration: 2 days

Objectives

At the end of this Section, participants will be able to:

- Understand how DHIS 2 supports data quality
- Use DHIS 2 as a data quality monitoring tool

Teaching Methods

- Lecture
- Group discussion
- Hands-on exercise
- Exercise

5.1. SECTION INTRODUCTION

DHIS2 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 DHIS2 design principles contribute to improving data quality, such as the idea of harmonizing 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 harmonized 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:

5.2. DATA INPUT VALIDATION

The most basic way of data quality check in DHIS2 is to make sure that the data being captured is on the correct format. The DHIS2 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 DHIS2 are explained in the user manual in the chapter on data elements.

5.3. MIN AND MAX RANGES

To stop typing mistakes during data entry (e.g typing '1000' instead of '100') the DHIS2 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 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.

5.4. 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 organization 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.

5.5. 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.

5.6. COMPLETENESS AND TIMELINESS REPORTS

Completeness reports will show how many data sets (forms) that have been submitted by organization 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 organization 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.

Annexes 4: RDQA Tool

Part 1 Data verification						
A - Documentation Review:		Indicator 1	Indicator 2	Indicator 3	Indicator 4	Comments
Review availability and completeness of all indicator source documents for the selected reporting period.						
1	Review available data sources for the reporting period being verified. Are all necessary data sources available for review?					1. 2. 3. 4.
	If no, determine how this might have affected reported numbers.					
2	Are all available data sources complete?					1. 2 3. 4.
	If no, determine how this might have affected reported numbers.					
3	Review the dates on the data sources. Do all dates fall within the reporting period?					1. 2. 3. 4.
	If no, determine how this might have affected reported numbers.					