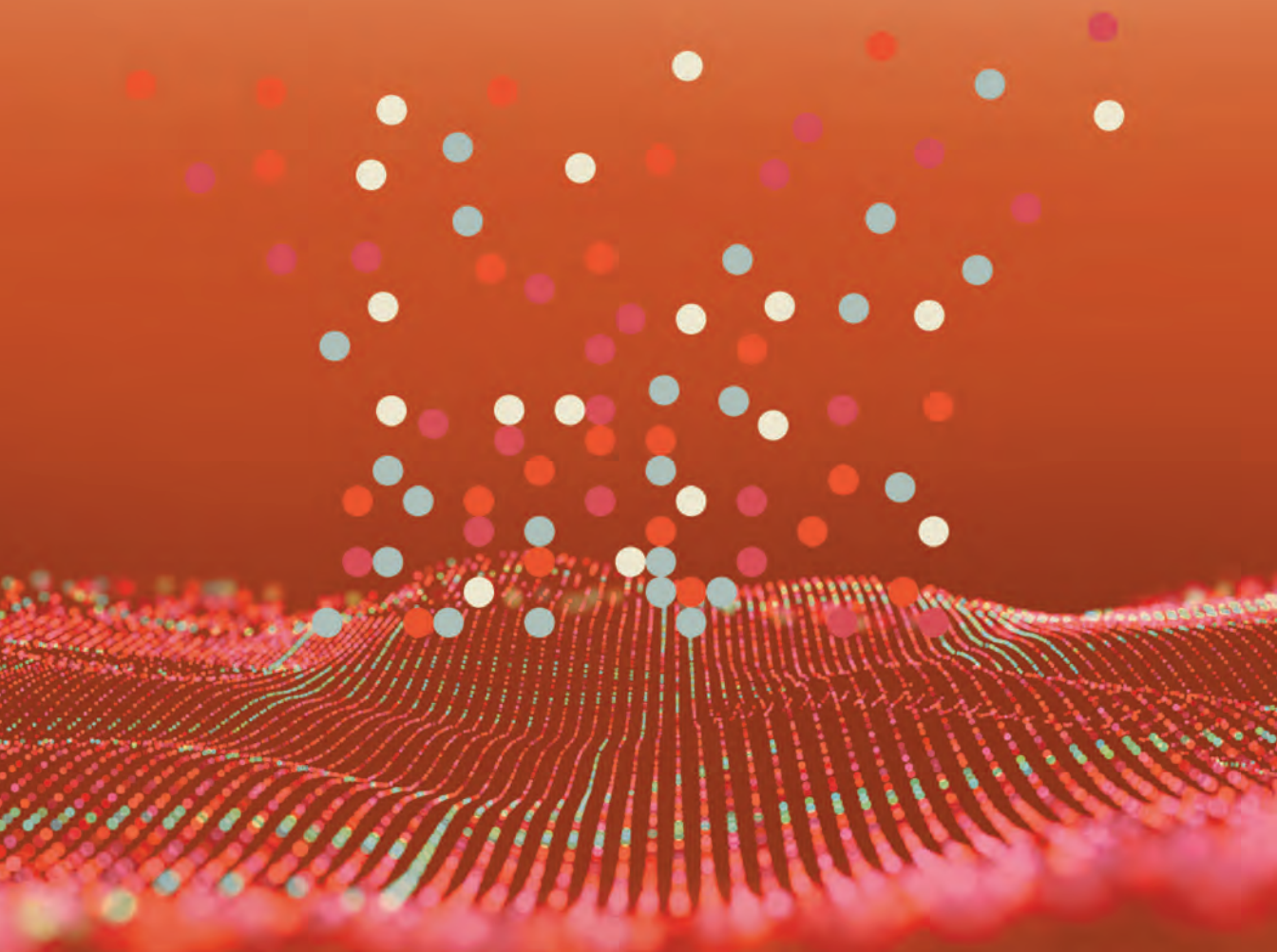
ICMR | NIMS
INDIAN COUNCIL OF MEDICAL RESEARCH | NATIONAL INSTITUTE OF MEDICAL STATISTICS

NDQF
National Data Quality Forum

National Guidelines for DATA QUALITY in Surveys



About ICMR-National Institute of Medical Statistics (ICMR-NIMS)

ICMR-NIMS is the premier medical statistics institute under the aegis of the Indian Council of Medical Research (ICMR), New Delhi.

The mandate of the institute is to complement the medical, health, epidemiology and surveillance, biomedical and behavioural research with development, application, and dissemination techniques in the statistical arena throughout India. This is achieved through statistical modelling, big data analysis and programme evaluation, technical and statistical consultation, education, and training as well as research and advocacy. ICMR-NIMS is designated by the National AIDS Control Organisation as the national institute since 2003 for estimation of HIV burden in India and the states/Union Territories. It also hosts the Clinical Trials Registry-India (CTRI) intending to encourage all clinical trials conducted in India to be prospectively registered. The institute's professional, well-qualified and experienced teams support government, researchers, and academia in advancing research and promoting sound statistical practices to inform public policy. ICMR-NIMS also offers trained human resource with the capability to carry out interdisciplinary research in the field of medical statistics.

National Guidelines for **Data Quality** in Surveys

July 2021

ICMR-National Institute of Medical Statistics, New Delhi

Suggested citation: ICMR-National Institute of Medical Statistics. 2021. National Guidelines for Data Quality in Surveys. New Delhi: ICMR-NIMS

ISBN: 978-93-5635-982-6

© ICMR-National Institute of Medical Statistics

The use of content from this guideline is encouraged with due acknowledgement to ICMR-National Institute of Medical Statistics. No part of the publication can be reproduced without the prior permission of ICMR-NIMS, New Delhi, India.

Contact: Director, ICMR-National Institute of Medical Statistics
nims.director@icmr.gov.in

CONTENTS

Messages
Preface
Acknowledgements
About National Data Quality Forum
Abbreviations
Preamble
Scope of the Document

1. Data Quality Monitoring 1-17
   - 1.1 General Principles of Data Quality Assurance in Surveys 2-4
   - 1.2 Data Quality Framework for Surveys 5-9
   - 1.3 Sources of Errors and Biases in Surveys 10-11
   - 1.4 Data Quality Assurance - Management Plan and Teams 12-13
   - 1.5 Procedures for Monitoring the Management Plan Implementation 14
   - 1.6 Quality Criteria for Reviewing Survey Protocols 15
   - 1.7 Ethics and Data Quality 16-17

2. Quality Assurance During the Preparatory Phase 19-35
   - 2.1 Study Design - Quality Assurance Assessment and Guidance 20
   - 2.2 Sampling Design - Quality Assurance 21-22
   - 2.3 Survey Tools 23-25
   - 2.4 Quality Considerations in Designing Data Entry Applications 26-27
   - 2.5 Quality Considerations When Recruiting Survey Investigators 28
   - 2.6 Training 29-30
   - 2.7 Preparatory Steps for Quality Assurance of Anthropometric Measurements 31-32
   - 2.8 Preparatory Steps for Quality Assurance of Biological Sample Collection 33-35

3. Implementation of Quality Assurance Activities During the Data Collection Phase 37-63

3.1 Summary of Considerations for Quality Assurance During Data Collection 39

3.2 Steps for Monitoring Survey Data Collection Quality 40-42

3.3 Steps for Monitoring of Anthropometric Data Quality 43-44

3.4 Steps for Monitoring Biomarker Sample Collection, Storage and Transportation Processes 45-50

3.5 Mechanism for Coordination Within and Between the Quality Assurance Team and the Main Survey Team 51

3.6 Tools to Monitor Quality of Field Data Collection 52-53

3.7 Use of Paradata to Improve Data Quality 54-55

3.8 Type of Analytics on Paradata to Present Data Quality Metrics 56-57

3.9 Using Dashboards to Monitor Data Quality 58-59

3.10 Indicators to Measure Data Quality for Providing Feedback to Investigators During Data Collection 60-61

3.11 Documentation of Data Quality Assurance 62-63

4. Data Quality Assessments Post Data Collection 65-74

4.1 Post Survey: Profiling Survey Data 66-70

4.2 Sample Weights and Sampling Errors 71-72

4.3 Data Quality Metrics: Calculation of Non-Sampling Errors 73-74

5. Use of Machine Learning Techniques in Improving Data Quality 75-78

References 79-82

Selected Definitions 83-86

List of Contributors 87

डॉ. विनोद कुमार पॉल
सदस्य
Dr. Vinod K. Paul
MEMBER

सत्यमेव जयते

भारत सरकार
नीति आयोग, संसद मार्ग
नई दिल्ली–110 001
Government of India
NATIONAL INSTITUTION FOR TRANSFORMING INDIA
NITI Aayog, Parliament Street
New Delhi-110 001
Tele. : 23096809 Fax : 23096810
E-mail : vinodk.paul@gov.in

19th July, 2021

Message

In today's world, major policy decisions and programme planning are data driven. Credibility, authenticity and quality of data are critical for making the data affable to larger translational and programme design endeavours. In India, there are ample data available majorly through surveys. Although each survey has theirown data quality checks and balances, a wholistic implementation of data quality assurance is often missing. There is a need to choose the wholistic approach to data-quality while conducting surveys – from planning to reporting of findings. Importantly, for doing so, adoption of uniform data-quality guidelines by each of these surveys is quintessential.

Being at the forefront of India's medical and health research, the Indian Council of Medical Research (ICMR) has always been emphasising on production of quality data. I am pleased to see that in the much-needed endeavour of National Data Quality Forum (NDQF) under the aegis, ICMR–National Institute of Medical Statistics (ICMR-NIMS), has developed a valuable document "**National Guidelines for Data Quality in Surveys".** I am sure, these guidelines present the best and most recommended practices for generating high-quality data and laud the efforts of the NDQF.

I congratulate ICMR, especially ICMR-NIMS for this timely and useful resource.

(Vinod Paul)

स्वच्छ भारत
एक कदम स्वच्छता की ओर

सत्यमेव जयते

प्रोफेसर (डा.) बलराम भार्गव, पदम श्री
एमडी, डीएम, एफआरसीपी (जी.), एफआरसीपी (ई.), एफएसीसी,
एफएएचए, एफएएमएस, एफएनएएस, एफएएससी, एफ.एन.ए., डी.एस.सी.
सचिव, भारत सरकार
स्वास्थ्य अनुसंधान विभाग
स्वास्थ्य एवं परिवार कल्याण मंत्रालय एवं
महानिदेशक, आई सी एम आर

Prof. (Dr.) Balram Bhargava, Padma Shri
MD, DM, FRCP (Glasg.), FRCP (Edin.),
FACC, FAHA, FAMS, FNASc, FASc, FNA, DSc
Secretary to the Government of India
Department of Health Research
Ministry of Health & Family Welfare &
Director-General, ICMR

भारतीय आयुर्विज्ञान अनुसंधान परिषद
स्वास्थ्य अनुसंधान विभाग
स्वास्थ्य एवं परिवार कल्याण मंत्रालय
भारत सरकार
वी. रामलिंगस्वामी भवन, अंसारी नगर
नई दिल्ली - 110 029

Indian Council of Medical Research
Department of Health Research
Ministry of Health & Family Welfare
Government of India
V. Ramalingaswami Bhawan, Ansari Nagar
New Delhi - 110 029

MESSAGE

India is endowed with rich data generated through multiple sources that guide our national programmes and policies. However, these data often suffer quality challenges for different reasons. There is an impending need for data-quality guidelines for demographic, health and nutrition surveys to improve the quality of the data that promotes evidence-based decision making.

Indian Council of Medical Research (ICMR) encourages the production of high-quality data and implement quality measures at each stage of survey viz., design, training, data collection, processing and dissemination. In this endeavour, I am happy that ICMR-National Institute of Medical Statistics (ICMR-NIMS) has come out with a valuable document titled **"National Guidelines for Data Quality in Surveys".** I hope that these guidelines will become best practices across the board for generating high-quality data. Of course, such an endeavour requires careful coordination and continued monitoring to achieve the desired result.

My hearty compliments and accolades for the perseverance and dedication of the National Data Quality Forum (NDQF) team for bringing out this publication. I am sure that the guidelines will be useful to all stakeholders engaged in conducting surveys and look forward to their implementation for ensuring data quality.

Balram Bhargava

(Balram Bhargava)

Tele.: 26588204, 26589620, Fax (Off.) : 91-11-26588662, E-mail: secy-dg@icmr.gov.in

आईसीएमआर–राष्ट्रीय आयुर्विज्ञान सांख्यिकी संस्थान
(भारतीय आयुर्विज्ञान अनुसंधान परिषद)
स्वास्थ्य अनुसंधान विभाग, स्वास्थ्य एवं परिवार
कल्याण मत्रालय, भारत सरकार
अंसारी नगर, नई दिल्ली – 110029

ICMR - NATIONAL INSTITUTE OF MEDICAL STATISTICS
(INDIAN COUNCIL OF MEDICAL RESEARCH)
Department of Health Research, Ministry of Health
and Family Welfare, Government of India
Ansari Nagar, New Delhi - 110029

Phone : 91-11-26588803
Telefax : 91-11-26589635
Email : nims.director@icmr.gov.in
: dr_vishnurao@yahoo.com

डॉ. एम. विष्णु वर्धना राव
एमएससी (स्टेटिस्टिक्स), एमटेक (आईटी), पीएचडी (स्टेटिस्टिक्स)
निदेशक
Dr. M. Vishnu Vardhana Rao
M.Sc(Stat), M.Tech(IT), PhD(Stat)
Director

Preface

Each year various public and private agencies collect data on multiple dimensions of policy and programme planning. High-quality data is the lifeblood of functional and responsive health systems. Reliable, complete and timely information is vital to shape policies and to ensure progress towards Sustainable Development Goals (SDGs). There is a general perception that these data often suffer quality challenges for various reasons. In recent years, data quality improvement efforts have been undertaken in both public and private sectors at independent/institutional levels. Convergence of these efforts is yet to be achieved in the form of an integrated platform at the national level to improve data quality. To address these issues, NDQF was launched in 2019 with the vision of improving quality of data that promotes evidence-based decision-making.

NDQF has since come out with its maiden publication 'National Guidelines for Data Quality in Surveys'. It provides a set of protocols for the survey with an overall aim to improve and maintain data quality in varied settings as per need of the users. The purpose includes outlining the functions of data quality assurance system that maintains transparency, credibility, accuracy and consistency. It aims to provide framework, identify data quality issues and ease out solutions. These guidelines aid in the application of data analysis techniques, which are valuable tools in monitoring the data quality assurance activities.

The guidelines deliberate on issues like design, method, monitoring, data profiling and technology for maintaining data quality. It encompasses assessing the data quality in the areas pertaining to: 1) Study design; 2) Sampling design; 3) Survey errors; 4) Survey tools; 5) Demographic, health and anthropometry measurements, 6) Biomarkers; 7) Monitoring tools; 8) Paradata; 9) Machine learning. The trustworthiness of any system can be established if the data meets certain standards. The guidelines encourage the use of standard methodologies, tools, techniques and sharing of skills resulting in rapid changes in data acquisition methods and in upcoming research.

The document will be a useful guide for researchers and policy makers in planning and conducting demographic, epidemiological, nutrition and allied surveys.

(M. Vishnu Vardhana Rao)

Acknowledgements

The document entitled 'National Guidelines for Data Quality in Surveys' is an outcome of joint efforts, hard work and involvement of many national and international experts representing varied disciplines, departments and organisations. I would like to thank everyone who contributed for its successful preparation.

I would like to express my profound sense of reverence and gratitude to Prof. (Dr.) Balram Bhargava, Secretary, Department of Health Research and Director General, Indian Council of Medical Research, New Delhi, as Steering Committee Chair for his guidance, supervision and leadership in the formulation and finalisation of the guidelines.

The guidance and expertise extended by other members of the Steering Committee- Dr. Pronab Sen - Director, IGC India, New Delhi and Ex-Secretary, MoSPI and Chief Statistician of India; Dr. Nivedita Gupta - Chief Director (Stats) MoHFW, New Delhi; Shri. Shankar Lal Menaria - Addl. Director General, National Statistics Office, MoSPI, New Delhi; Dr. P. Ashok Babu - Director, POSHAN Abhiyaan, Anganwadi Services (ICDS), MWCD, New Delhi; Dr. Shekhar Shah - Director, NCAER, New Delhi; Dr. T. K. Roy - Ex-Director, International Institute for Population Sciences (IIPS), Mumbai; Shri. Pankaj Shreyaskar - Director, Survey Coordination Division, National Statistics Office, MoSPI, New Delhi; Dr. K. S. James - Director and Senior Professor, International Institute for Population Sciences, Mumbai; Dr. Suneeta Krishnan - Country Lead (MLE), Bill & Melinda Gates Foundation, New Delhi; Shri. Sanjeev Kumar - Deputy Director General, Office of the Registrar General of India, New Delhi; and Mr. D. K. Ojha - DDG (Stats, HMIS), MoHFW, New Delhi, at all stages of the preparation of the guidelines is gratefully acknowledged.

I sincerely acknowledge the contributions of the Technical Advisory Committee consisting of Dr. Pronab Sen - Director, IGC India, New Delhi and Ex- Secretary, MoSPI and Chief Statistician of India; Dr. Sanjay Mehendale - Ex-Addl. DG, ICMR, New Delhi; Dr. J. K. Banthia - Ex-RGI & Census Commissionner of India, New Delhi; Dr. Padam Singh - Ex-Addl. DG, ICMR, New Delhi; Dr. Arvind Pandey - Ex-Director, ICMR-NIMS, New Delhi; Dr. Bontha V. Babu - Head, SBR Division, ICMR, New Delhi; Dr. Rakhi Dandona - Professor, Public Health Foundation of India, New Delhi; Dr. Dileep Mavalankar - Director, Indian Institute of Public Health, Gandhinagar; Dr. P. Kumaraguru - Professor, Indraprastha Institute of Information Technology, New Delhi; Dr. Tavpritesh Sethi – Associate Professor, Indraprastha Institute of Information

Technology, New Delhi; Mr. Robert Johnston - Nutrition Specialist, UNICEF, New Delhi; Dr. Sharon Buteau - Executive Director, IFMR-LEAD, New Delhi; Dr. Rinku Murgai - Lead Economist, World Bank, New Delhi; Dr. Divya Nair - IDinsight, New Delhi; Dr. S. Sridhar - Technical Director, Bihar TSU, CARE India, Patna; Dr. S. Pyne - Professor, Pittsburgh University, Pittsburgh; Ms. Priyanka Dutt - Senior Research Manager, BBC Media Action Trust, New Delhi; and Dr. Yujwal Raj, Independent Consultant, Hyderabad, for its guidance and advice provided in the formulation of the guidelines.

I acknowledge wholeheartedly the sincerity, belongingness and commitment of the members of the Peer Review Group - Dr. T. K. Roy - Ex-Director, IIPS, Mumbai; Dr. Fred Arnold - Senior Fellow, ICF, Rockville; Dr. G. Bhanuprakash Reddy – Scientist G, ICMR-NIN, Hyderabad; Dr. H. P. S. Sachdev - Paediatric Consultant, Sitaram Bhartia Hospital, New Delhi; and Dr. U. V. Somyajulu - Sigma Research Ltd., New Delhi - for critically reviewing the draft guidelines and providing invaluable feedback.

The publication of this document would not have been achieved without the commitment and contributions of a team of scientists from ICMR-National Institute of Medical Statistics (NIMS), New Delhi, researchers of the Population Council, New Delhi; and NDQF project staff. I express my gratitude for their persistent and unstinted support in bringing out this document in short time frame. I would also like to thank Mr. Vaibhav Malhotra, Project Officer, NDQF at the ICMR-NIMS, and Ms. Radhika Dhingra – Assistant Program Officer and Ms. Ramandeep Kaur – Senior Executive Assistant at the Population Council, for their invaluable support at different stages of preparation and production of this document.

Last, but not the least, I am extremely grateful to the Bill & Melinda Gates Foundation for its technical and financial support through the Population Council. I am especially thankful to the Population Council for its facilitation right from initiation, till the completion of the document.

Director
ICMR-NIMS, New Delhi

About National Data Quality Forum

National Data Quality Forum (NDQF), launched in July 2019, is a multi-stakeholder, collaborative platform housed at the Indian Council of Medical Research (ICMR) - National Institute of Medical Statistics (NIMS) and co-led by the Population Council, New Delhi; supported by the Bill & Melinda Gates Foundation (BMGF). The forum extends partnership with government institutions, agencies, and ministries for identifying opportunities to build systems for ensuring data quality, further equipping institutions with solutions and strategies to contribute to data quality improvement. Network institutions of NDQF include, but are not limited to, Registrar General of India (Ministry of Home Affairs), International Institute for Population Sciences (IIPS), National Statistics Office (NSO, Ministry of Statistics and Programme Implementation), Statistics Division (MoHFW), Indraprastha Institute of Information Technology-Delhi (IIIT, D), Indian Statistical Institute-Delhi (ISI, Delhi), Jawaharlal Nehru University-Delhi, Centre of Development Studies-Thiruvananthapuram, and Ministry of Women and Child Development (MWCD).

The NDQF envisions providing a framework for advanced data quality monitoring, process audits and analytics and building the capacity of data collectors for improving the quality of surveys and administrative data. The goals of NDQF are:

- Improve health and demographic data ecosystem in India
- Strengthen data quality as well as the systems that create and manage data
- Deepen interests in the importance of good quality data among both producers and users of data
- Educate consumers to demand data of good quality

NDQF's approaches include bringing together producers and users of data on a single platform, facilitating knowledge exchange and equipping them with appropriate tools and guidelines for data quality assurance; generating novel solutions for improving data quality by facilitating development and testing of innovations; developing national guidelines on data quality assurance; build capacity within data-producing institutions to implement the data quality assurance strategy co-developed with NDQF and helping in institutionalisation of strategy.

Abbreviations

ASHA	Accredited Social Health Activist
BMGF	Bill & Melinda Gates Foundation
BOND	Biomarkers of Nutrition for Development
CAB	Clinical, Anthropometric, and Biochemical
CAPI	Computer Assisted Personal Interviewing
CGHS	Central Government Health Scheme
CNN	Convolutional Neural Network
CNNS	Comprehensive National Nutrition Survey
CSR	Corporate Social Responsibility
CTRI	Clinical Trials Registry-India
CV	Coefficient of Variation
DHS	Demographic and Health Survey
DLHS	District Level Household Survey
DQA	Data Quality Assurance
DQMS	Data Quality Management Structure
EURECCA	European Micronutrient Recommendations Aligned
FCT	Field Check Tables
FMC	Field Monitoring Checklist
GATS	Global Adult Tobacco Survey
GPS	Global Positioning System
HIV	Human Immunodeficiency Virus
ICMR	Indian Council of Medical Research
IIPS	International Institute for Population Sciences

LASI	Longitudinal Ageing Study in India
ML	Machine Learning
MUAC	Mid Upper Arm Circumference
MWCD	Ministry of Women and Child Development
NDQF	National Data Quality Forum
NFHS	National Family Health Survey
NHANES	National Health and Nutrition Examination Survey
NIMS	National Institute of Medical Statistics
NLP	Natural Language Processing
NSO	National Statistics Office
OCR	Optical Character Recognition
ODK	Open Data Kit
PSU	Primary Sampling Unit
QA	Quality Assurance
QC	Quality Control
RCH	Reproductive and Child Health
RNN	Replicator Neural Networks
SE	Standard Error
SOP	Standard Operating Procedure
SRS	Simple Random Sample
SVM	Support Vector Machine
TOT	Training of Trainers
TRF	Test Requisition Form

Preamble

Aim

To generate quality survey data by mitigating errors and biases that may creep in during survey design, data collection and analysis.

Objectives

- To bring awareness among producers and users about best practices around data quality
- To assist in evolving strategies for data quality assurance mechanisms for surveys
- To guide institutionalisation of data quality assurance mechanisms

Targeted Stakeholders

These guidelines serve various categories of data producers and users.

Data Producers	Data Users
• Research organisations • Academic institutions • Individual researchers • Ministries/Government departments • National and international agencies • Commercial survey agencies • Non-profit organisations • Development partners	• Programme implementers • Programme managers • Policy makers • Academic institutions • Individual researchers • Development professionals • Civil society organisations • Corporate Social Responsibility (CSR) foundations/organisations • Donor agencies

Scope of the Document

The purpose of this guideline document is to provide a comprehensive list of guiding principles and best practices in data quality of all sample surveys with specific reference to demographic, nutrition and health surveys.

This document provides insight on crucial steps that need to be followed right from the beginning to ensure data quality. The document is divided into three sections based parts on three broad phases of a survey:

- Preparatory phase
- Data collection phase
- Post data collection phase

It lists out the points that need to be borne in mind during the preparatory phase, including the study design, sampling, survey tools and manuals. It guides readers on key quality considerations while designing and developing a data entry package, quality assurance protocols that ensure quality of survey, anthropometry and biomarker data, recruitment and training of survey investigators, and assessment of trained health/research investigators.

The guideline document describes the quality assurance activities in the data collection phase, monitoring quality of data collection using multiple tools, suggests usage of data entry parameters and quality dashboard during a field survey.

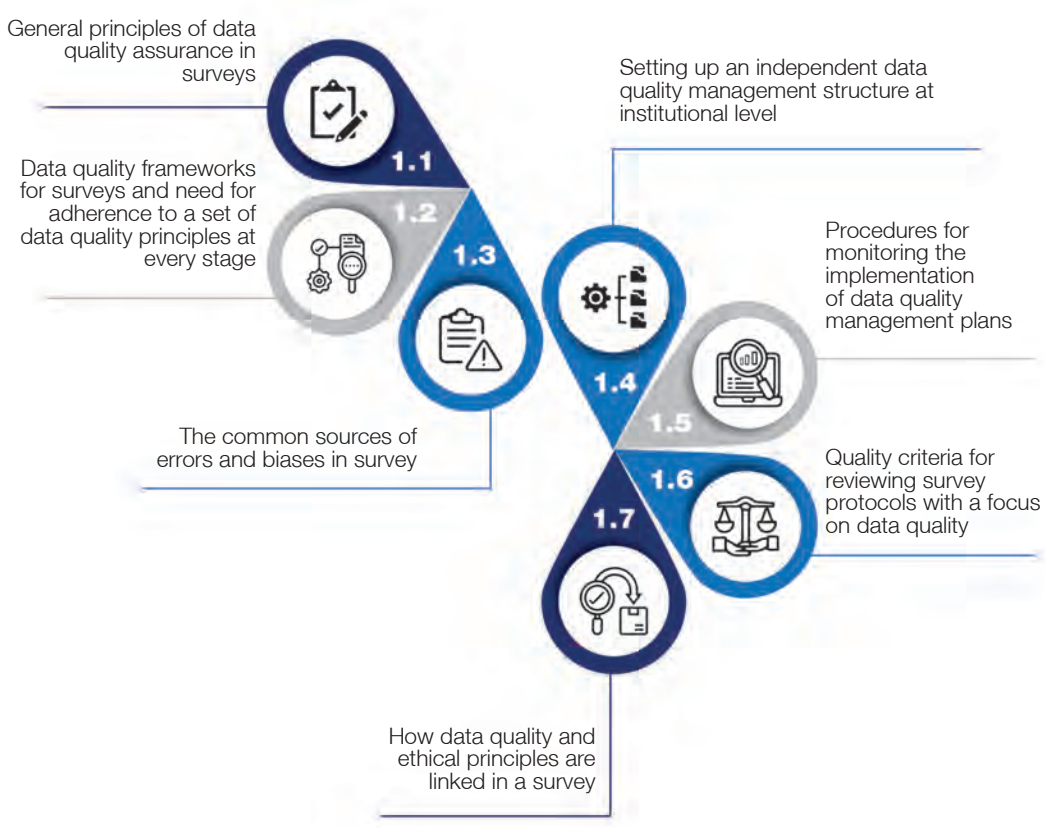
The document also guides on post data collection quality checks, reviewing the data and using appropriate data quality analytics. It outlines different techniques that can be employed for assessing quality of data, and estimation of sampling and non-sampling errors. Application of different machine learning techniques in the assessment of quality of the surveys is made easy with this document as it provides tips on the use of technology and checklists for quick guidance, wherever needed.

These guidelines are useful for a wide range of audiences viz., government and private data producers and users, national and regional level policy makers, and technical staff at the ministries and organisations, besides academic, research institutions and survey agencies. It is immensely helpful in guiding the planning, designing and execution of sample surveys and achieving high quality data.

However, it is important to note that this document is not a manual, it rather specifies the guiding principles to follow before, during and after a field survey for better data quality. These guidelines are meant for in-person quantitative surveys and not for those carried out through telephone or web. A separate set of guidelines need to be prepared for data quality in such surveys.

1. Data Quality Monitoring

A well-defined data quality assurance mechanism is a non-negotiable component of any survey that aims to collect high quality data. Setting up such a system requires understanding of the general principles of data quality assurance, its importance, establishing the institutional structure for implementation of the quality management plans, preparing quality criteria for reviewing survey protocol and plans for adhering to ethical principles. Below are some key steps for developing a data quality assurance mechanism for surveys.



1.1 General Principles of Data Quality Assurance in Surveys

Building a quality assurance mechanism is key to data quality in surveys. The efforts towards this start with survey planning. The basic quality requirements are:

- **Developing a data quality management plan**
- **Implementing data quality assurance activities**
- **Analysing data quality**
- **Providing feedback/report on data quality**

Every survey conducted in the fields of demography, health, and nutrition (or likewise) must prepare a quality assurance plan during the planning process. The quality assurance plan should list detailed steps and activities. To build such a plan, the survey implementing agencies or the survey coordinating organisations may consider:

1. **Building a data quality assurance team during the planning phase of the survey that will undertake tasks to ensure quality control at each step of the planning process, implementation, and analyses; these may include but are not limited to:**

    - Guiding and supporting data collection agencies/teams within the organisation on data quality procedures
    - Finding resources and matching them to the quality management plans and activities
    - Setting up support systems for data quality teams in the field
    - Coordinating quality assurance through each step of the survey process

2. **Building a comprehensive and sustainable data quality management plan that shall include the profiles and procedures for quality assurance based on the survey/study goals and objectives. These may include but are not limited to:**

    - Creating a flowchart of the management plan for quality assurance

- Describing the structure of the team for quality assurance during survey implementation
- Creating tools that help measure data quality on an ongoing basis
- Developing procedures for both quality management plan and quality assurance procedures/activities in the field

3. Establishing procedures that need to be implemented from initiation to completion of the study to ensure data quality; critical elements of quality assurance activities and their implementation are:

a. During the preparatory phase

- Examine the appropriateness of the study design for the objectives of the study.
- Assess the sampling methodology and its appropriateness to make it a representative or generalisable study for the population it intends to cover.
- Review different dimensions of the survey tool to eliminate bias.
- Ensure that quality assurance steps are listed for recruitment, training and retention of survey investigator teams.
- Ensure quality assurance steps are listed for anthropometry and biomarker data collection (if included in the study).
- Assess and certify trained health/research investigators.
- Standardise survey procedures.

b. During the data collection phase

- Perform quality assurance activities through planned field visits.
- Monitor quality of data collection intensely and regularly.
- Measure quality of data using tools designed under the quality management plan.
- Analyse and prepare reports on quality assurance on and off the field.
- Take steps based on the observed quality of data collection, including but not limited to retraining field investigators, modifying field plans if required, and strengthening monitoring systems.

- Update quality assurance activities as the survey implementation matures in field.
- Document all the steps taken, modifications made to the survey or quality assurance during the data collection phase.

c. Post data collection phase

- Perform analytics on paradata to monitor the progress in data quality.
- Perform analytics on key study indicators to examine investigators' bias or indicative patterns.
- Present dashboards on data consistency.
- Report data quality measures, including but not limited to sampling error, non-sampling error and investigator bias.

4. Providing constructive feedback on data quality throughout the survey process is critical to achieving good quality data. Feedback needs to be timely, pointed towards identified issues and offer solutions. Feedback can be given using digital platforms or via face-to-face meetings based on evidence generated from paradata, metadata or study data.

1.2 Data Quality Framework for Surveys

Maintaining data quality in survey requires adhering to quality principles and protocols at each stage of the survey. To ultimately improve survey data quality, assessing the quality dimensions of obtained survey data, documenting data quality measures and establishing mechanism for regular feedback from data users and producers are required [1].

The following diagram provides the specific set of principles at each stage of the survey.

Quality dimensions related to survey data and processes, associated attributes, metrics/indicators, and sections where these issues are dealt with are provided in the table below.

Dimensions	Attributes	Metrics/ Indicators	Sections Where These Issues are Dealt with
SURVEY INSTITUTION			
Professional Independence	Free of interference from policy, regulatory bodies or institution	Comprehensive legislation and/or code of professional ethics maintained by the institution	Data quality assurance: Management plan and teams (1.4, 1.5); data profiling (4.1); calculation of non-sampling errors/bias (4.3)
Commitment to Quality	Publicly available institutional policy statement and/ or processes to ensure quality in survey data	Existence of institutional body for data quality assurance; availability of quality guidelines	Data quality assurance: Management plan and teams (1.4, 1.5); study design (2.1); sampling design (2.2); survey tools (2.3); recruitment of investigators (2.5); training (2.6); monitoring of survey data (3.2-3.4)
Integrity	Institution follows values and practices that maintain user confidence in data	Document and appraise public about the processes and protocols followed and the statistical issues faced during the survey	Quality criteria for review of survey protocols/bids (1.6); ethics (1.7); documentation on data quality (3.11); data profiling (4.1)
Resource Sufficiency	Adequacy of human, financial and technical resources	Strategic planning process, recruitment of relevant staff, proper allocation and utilisation of financial and technical resources	Data quality assurance: Management plan and teams (1.4); quality criteria for review of survey protocols/bids (1.6); recruitment of investigators (2.5); training (2.6); monitoring of survey data (3.2-3.4)
Data Confidentiality	Policy to safeguard privacy of respondents, mechanisms to address security related to data storage and dissemination	Assuring and ascertaining data confidentiality using: • De-identification of respondents in the public dataset • Up-to-date data security measures	Quality criteria for review of survey protocols/bids (1.6); ethics (1.7)

Dimensions	Attributes	Metrics/ Indicators	Sections Where These Issues are Dealt with
SURVEY PROCESS			
Methodological Soundness	Appropriate study design, ethical procedures, sampling approaches, sample selection and study tools, interview approach, survey monitoring, and data recording	Engagement of right experts for various components of survey; adoption of internationally accepted standards for data collection and monitoring	Types of survey error (1.3); quality criteria for review of survey protocols/bids (1.6); study design (2.1); sampling design (2.2), survey tools (2.3); quality assurance of anthropometric and biological data (2.7, 2.8); calculation of sampling weight, sampling error (4.2) and non-sampling errors/bias (4.3)
Response Burden	Limit data collection tools to cover necessary aspects related to survey objectives	Well-designed data collection tools and use of sound scientific and technological procedures; pre-testing of tools to understand response burden	Study design (2.1); survey tools (2.3)
Cost-efficiency	Effective utilisation of available financial, technical and human resources	Use of modern technological and communication strategies for data collection and field monitoring	Quality criteria for review of survey protocols/bids (1.6); designing data entry application (2.4); coordination mechanism (3.5); tools to monitor data quality (3.6); use of paradata (3.7, 3.8); data quality dashboard (3.9); use of machine learning techniques (5)
SURVEY OUTPUT			
Relevance	Satisfy the data needs of the users	Tools used for the study cover the study objectives; indicators are standardised	Type of survey error (1.3); study design (2.1); sampling design (2.2); survey tools (2.3)
Accuracy	Closeness of estimate to reality	Assess coverage error, measurement error, non-response error, sampling error (CV, Variance, SE) of key estimates	Type of survey error (1.3); study design (2.1); sampling design (2.2); survey tools (2.3); data profiling (4.1); calculation of sampling weight, sampling error (4.2) and non-sampling errors/bias (4.3)

Dimensions	Attributes	Metrics/ Indicators	Sections Where These Issues are Dealt with
Reliability	Closeness of the initial estimate to the final one	Test-retest reliability, alternate-form reliability and internal consistency reliability of key estimates (Cronbach's Alpha, Kappa statistics)	Tools to monitor data quality (3.6); use of paradata (3.7, 3.8); data quality dashboard (3.9)
Accessibility/ Clarity	Data accessible to public with relevant metadata and data use guidelines	Open access data; clearly defined metadata; relevant documentation of data quality and data profiling	Ethics (1.7); data profiling (4.1); documentation on data quality (3.11)
Coherence	Consistency of different indicators within the same dataset; option of integration at various levels (for example, district, state, national, gender, age)	Tools contain questions to assess internal consistency; key geographic characteristics are provided in the unit level data	Survey tools (2.3); designing data entry application (2.4); data profiling (4.1)
Comparability	Allow comparability across geographies, time, domains	Standard design used in the survey and standardised indicators are available for comparison	Study design (2.1); survey tools (2.3)
Completeness	All data items recorded	Percent of missing values on different indicators	Data profiling (4.1)
Validity	Data collected following specific rules (for example, variables confirm to certain format, type, category and range)	Compare the data and metadata or documentation for all indicators	Data profiling (4.1)

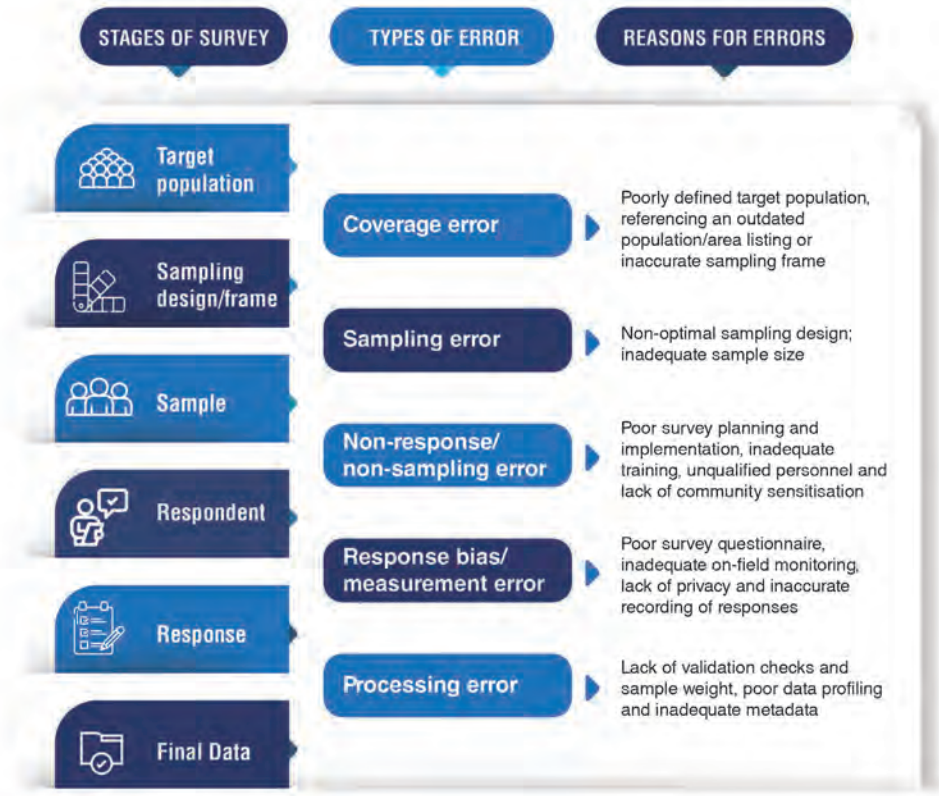
Note: The quality dimensions, quality attributes and metrics/indicators presented in this table were sourced from mutliple publications [2-9]

1.3 Sources of Errors and Biases in Surveys

Errors are an essential part of survey estimates. In a sample survey, errors occur because all the members of the sampling frame are not observed, rather estimates are drawn from a small representative segment of the target population [10-12]. Estimates from the sample population are used to infer on the characteristics of the target population, assuming that they are generalisable. The difference between the estimates drawn from the sample and the corresponding population parameter is termed as survey error. There are other types of errors that stem from inappropriate implementation of sampling design or problematic data collection, entry or processing. Further, some biases can occur making a population estimate unusable. Two types of survey errors are observed, some are measurable, and some are not:

Sampling Error: The deviation between a sample estimate and the population parameter under study, caused by sample selection, is generally referred to as the sampling error. These errors occur in the preparatory phase of a survey while conceptualising the sampling strategy. While sampling errors are inevitable in a survey, they can, however, be reduced by either increasing the size of the sample or by using stratification. Higher the sample size, closer will be the sample estimate to population value. If a population is heterogeneous, stratification can make the sample more representative of the population and thereby reduce error.

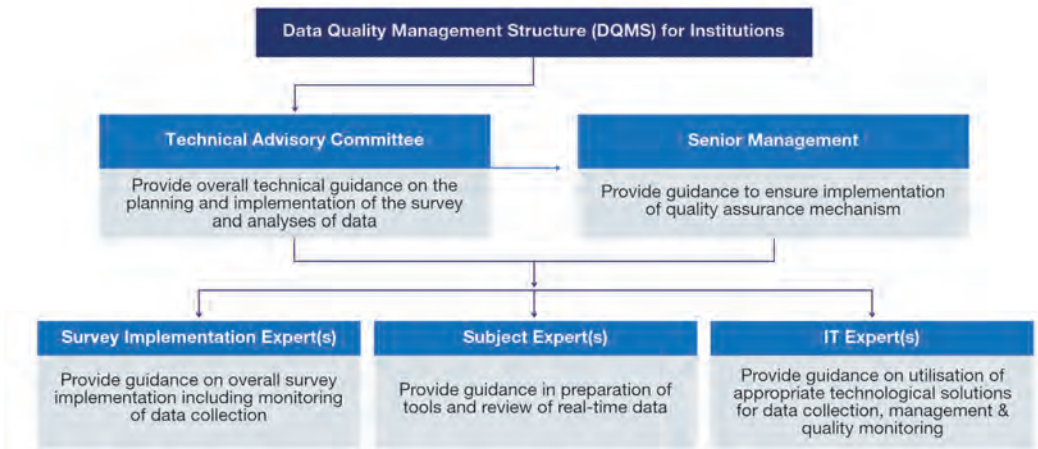
Non-Sampling Error: Errors occurring at various stages during data collection and the processing phase and not linked to the selection of sample are called non-sampling errors. Proper survey planning, robust training and standardisation of field investigators, standardised questioning can be instrumental in reducing non-sampling errors. Unlike sampling error, non-sampling errors cannot be directly estimated but can be reduced by ensuring the quality of the survey processes at each stage.

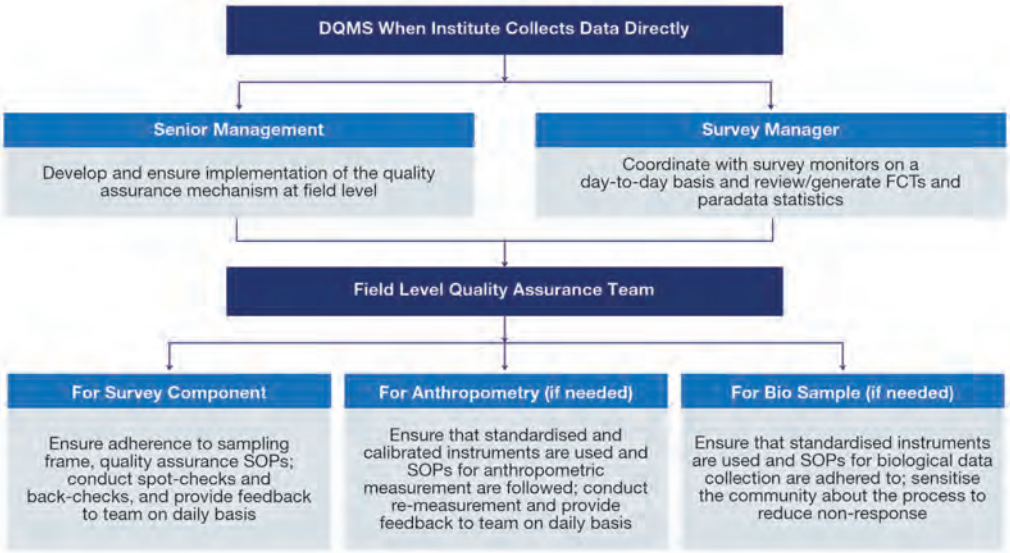


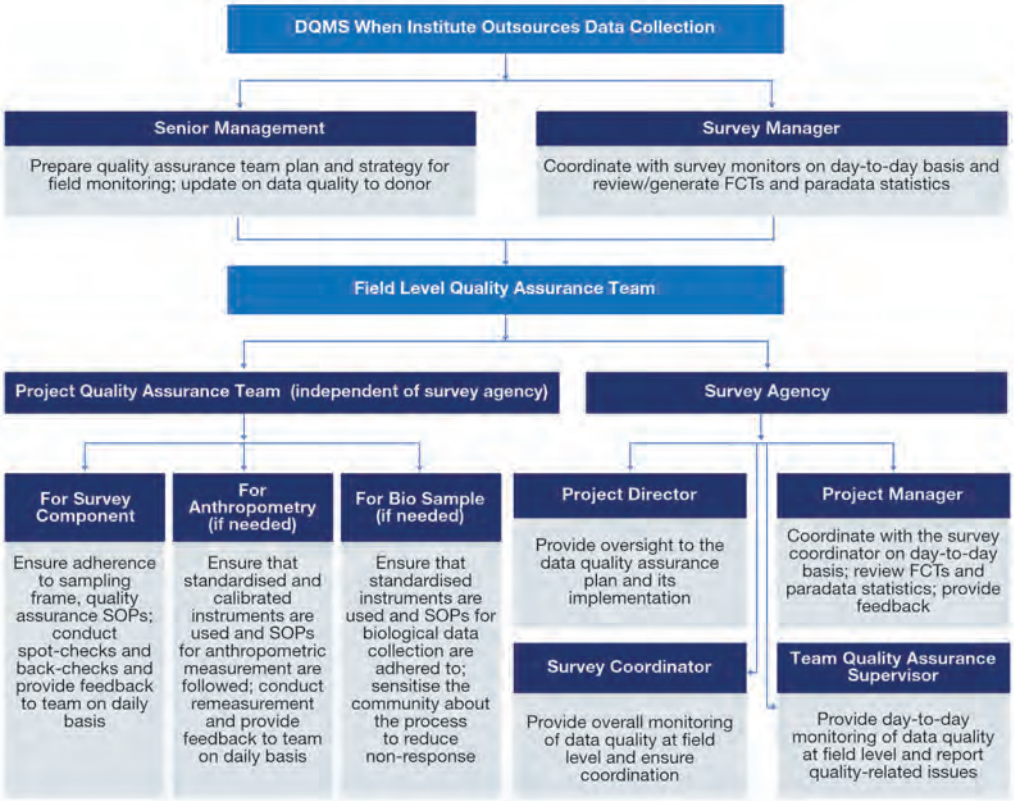
Biases in Survey: While errors in survey occur randomly, deviation of sample mean from the population mean can also be caused by different kinds of biases that occur systematically during the survey. For example, response bias is caused by a deliberate attempt to alter true response whereas coverage bias is caused by intentionally leaving out certain communities during the survey.

1.4 Data Quality Assurance – Management Plan and Teams

One of the ways data quality can be improved is through setting up an independent Data Quality Management Structure (DQMS) at the institution level. DQMS can be implemented at various levels depending on the overall objective of the survey and its implementation strategy. At the institution level, DQMS can provide overall guidance and strategic inputs to ensure the quality of data. A core team of experts can provide the required guidance to each component related to the survey data quality and monitor adherence to good practices. At the project level, depending on whether data collection is outsourced to a survey agency or is implemented by the institute/organisation/agency by itself, an appropriate data quality assurance team consisting of senior management and field level quality monitors can be constituted.




DQMS When Institute Collects Data Directly
Senior Management
Develop and ensure implementation of the quality assurance mechanism at field level
Survey Manager
Coordinate with survey monitors on a day-to-day basis and review/generate FCTs and paradata statistics
Field Level Quality Assurance Team
For Survey Component
Ensure adherence to sampling frame, quality assurance SOPs; conduct spot-checks and back-checks, and provide feedback to team on daily basis
For Anthropometry (if needed)
Ensure that standardised and calibrated instruments are used and SOPs for anthropometric measurement are followed; conduct re-measurement and provide feedback to team on daily basis
For Bio Sample (if needed)
Ensure that standardised instruments are used and SOPs for biological data collection are adhered to; sensitise the community about the process to reduce non-response


DQMS When Institute Outsources Data Collection
Senior Management
Prepare quality assurance team plan and strategy for field monitoring; update on data quality to donor
Survey Manager
Coordinate with survey monitors on day-to-day basis and review/generate FCTs and paradata statistics
Field Level Quality Assurance Team
Project Quality Assurance Team (independent of survey agency)
Survey Agency
For Survey Component
Ensure adherence to sampling frame, quality assurance SOPs; conduct spot-checks and back-checks and provide feedback to team on daily basis
For Anthropometry (if needed)
Ensure that standardised and calibrated instruments are used and SOPs for anthropometric measurement are followed; conduct remeasurement and provide feedback to team on daily basis
For Bio Sample (if needed)
Ensure that standardised instruments are used and SOPs for biological data collection are adhered to; sensitise the community about the process to reduce non-response
Project Director
Provide oversight to the data quality assurance plan and its implementation
Project Manager
Coordinate with the survey coordinator on day-to-day basis; review FCTs and paradata statistics; provide feedback
Survey Coordinator
Provide overall monitoring of data quality at field level and ensure coordination
Team Quality Assurance Supervisor
Provide day-to-day monitoring of data quality at field level and report quality-related issues

1.5 Procedures for Monitoring the Management Plan Implementation

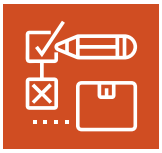
The monitoring process oversees all the tasks and metrics necessary to ensure that the protocols are implemented as planned, with specific reference to the data quality assurance team's scope, time and budget so that project risks are minimised.

This process involves comparing planned performance with actual performance, identifying risks and risk mitigation plans, determining appropriate action plans, and mapping personnel to accountability of action plans. This monitoring process is implemented throughout the life of the survey.

Key dimensions to monitor quality assurance management plan include, but not limited to:

- **Scope of work alignment/deviations**
- **Quality control steps executed as per the management plan**
- **Risks identified by quality assurance teams and steps taken to address them in partnership with survey teams**
- **Cost control**
- **Review of quality control indicators and actions taken**

The survey manager will have the responsibility to ensure proper implementation of the data quality management plan.



1.6 Quality Criteria for Reviewing Survey Protocols

Engagement of the right institution/agency (or) having a strong proposal is the prerequisite to good quality survey data. Before survey implementation, the agency/ institution must ensure inclusion of the following parameters in the proposal.

- **Experience of the team in conducting similar surveys**
- **Availability of an operational team with dedicated quality control staff at the institution level**
- **Plans for pre-testing of survey tools**
- **Survey implementation design and timeline**
- **Size and composition of data collection teams**
- **Necessary qualifications and experience of field staff that meet the needs of the survey**
- **Necessary qualifications and experience of master trainers**
- **Plans for piloting the survey implementation**
- **Plans for real-time data entry (for example, computer assisted personal interviewing) with data checks, if any**
- **Plans for back-checks and spot-checks of data collection (a specific percent of sample)**
- **Plans for generation of quality parameters, tracking quality of data by investigators and teams, and feedback mechanisms for the teams**
- **Possible data quality challenges/risks and potential mitigation steps**

If the data producing organisation (nodal agency) is inviting proposals to undertake the survey, the quality criteria (based on the above parameters) shall be integrated within the request for proposal and also be used for evaluation. For selecting bids, data quality assurance plan should have a considerable weightage in the bid-evaluation methodology.

1.7 Ethics and Data Quality

Compliance to research ethics helps ensure data quality and credibility of research. Ethical principles to be followed in a survey have been laid out in the available literature and guidelines. These ethical principles serve as a guide through planning, funding, and conducting research, as well as for data storing, analysis, sharing of data, use of data and dissemination of research findings. While it is imperative for any survey to seek approval from an Institutional Review Board (IRB) and to conform to standard operating guidelines and procedures of survey research ethics throughout the life cycle of the survey, the following key measures should be practiced for quality control and monitoring of implementation of ethical guidelines [13-18]:

1. **Ensure the survey uses participant information sheets and informed consents that adhere to national/international ethical guidelines. Additionally, for an ethically appropriate data quality monitoring system, a statement seeking participant consent for a follow up visit by field supervisors and other data quality monitors should be added to the informed consent. The purposes of these monitoring visits are to assess adherence of ethical protocols in the field and quality of information elicited during the interviews.**

2. **Design and implement a quality monitoring checklist to ensure that ethical guidelines are being followed during the informed consent process.**

3. **Organise training of staff engaged in quality assurance and field monitoring on ethical principles and steps to be taken while monitoring the field work or back checking the interviews including confrontation on previously collected information from the same respondent. Reiterate on principles including the use of social media by research staff to protect privacy and confidentiality of respondents.**

4. De-identify respondent's information from the publicly available dataset.

5. Store data in a secured platform with defined access.

6. Regular monitoring by the Institutional Review Boards to ensure adherence to the ethical protocols.

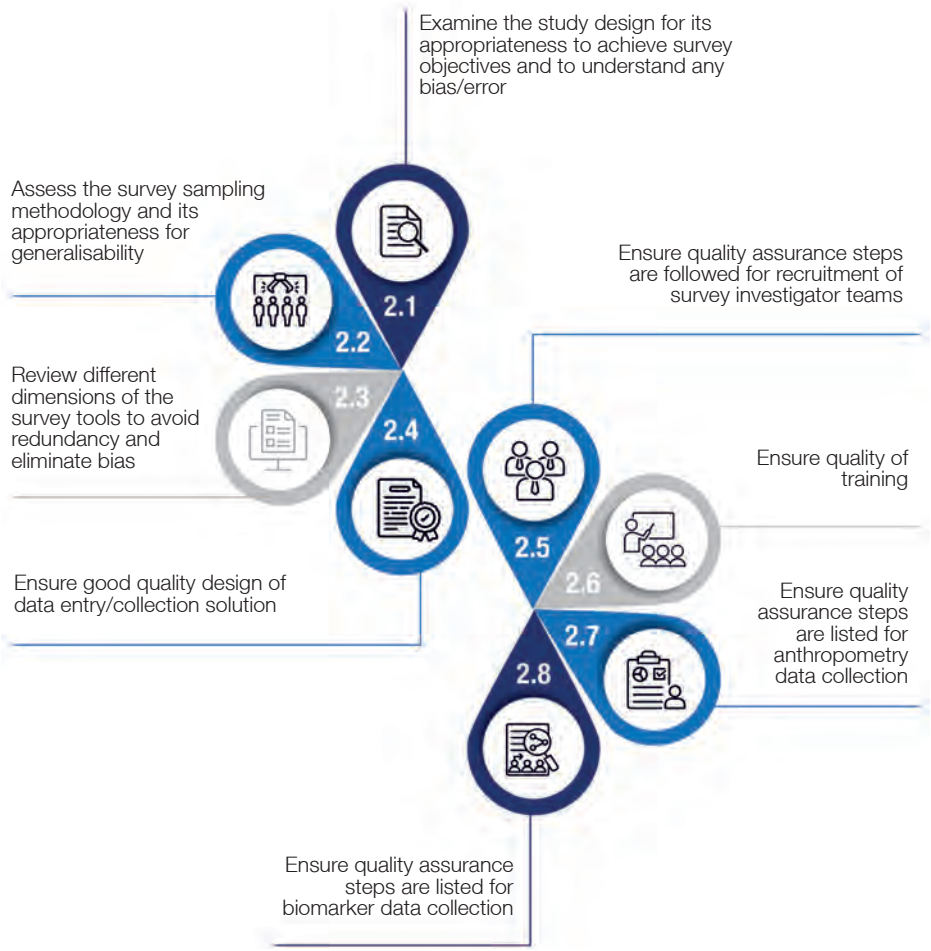
Technology Tip:

- Consent can be administered via using handheld devices through digital signature, voice, or video. For example, options for capturing digital signatures are available in data entry software like ODK, Survey CTO, KoBo Toolbox

This page has been intentionally left blank

2. Quality Assurance During the Preparatory Phase

The preparatory phase forms the foundation in quality control assurance for the whole survey. This section of the document outlines 'what' quality assurance parameters can be considered during the preparatory phase of the survey and 'how' they can be implemented. It lists out the quality assurance steps in developing study design, sampling, survey tools and manuals, recruitment of survey investigators, designing of data entry application, and training of survey investigators. Below are some of the key principles that can be followed during the preparatory phase of the survey:



2.1 Study Design – Quality Assurance Assessment and Guidance

There are a number of study designs available that can be adapted according to the objectives of the study. Broadly, these study designs can be categorised as observational and experimental. Details of different study designs can be found elsewhere [19].

The following parameters can help determine the quality of the study design:

1. **The study has unambiguous research questions or points of investigation.**
2. **The proposed study methodology is appropriate for the research questions.**
3. **Strategy for recruiting respondents is in accordance with the study's aim and there is no selection bias.**
4. **Data collection methods appropriately address the research questions.**
5. **Ethical considerations are accounted for in the study.**
6. **Rigorous data quality and management methods are described.**
7. **Strategy to address non-response during data collection is stated.**
8. **Steps for data analysis are clearly defined.**
9. **Steps for data utilisation are clearly articulated.**

Irrespective of the study design, the guidance provided in this document applies to any survey-based data collection.

2.2 Sampling Design – Quality Assurance

Adopting a proper sampling design is one of the most important steps in ensuring quality of survey estimates. Details on various sampling methods are available elsewhere [20-22]. Following are the principles for a good quality sampling design:

1. **Decide on an appropriate sampling design that is scientific as well as cost-effective. In population-based surveys, having more than one stage of selection often saves money.**

2. **Identify a proper sampling frame. If not readily available, create one through a listing exercise (for example, listing of households) or from registers (for example, ASHA registers of pregnant women).**

3. **Check the sampling frame for any exclusion or duplication of target sample units. If found defective, correct the frame before sample selection.**

4. **If information on variables that affect the outcome of interest is available in the sampling frame, use them to stratify the frame, for example, female literacy, ethnicity, occupation. This helps reduce sampling error and improve design efficiency.**

5. **Wherever possible, stick to EPSEM (Equal Probability of Selection Method) design.**

Checklist

- Sampling design documented ☐
- Adequate sample size ☐
- Appropriate sampling method chosen ☐
- Stratification used (if applicable) ☐
- Accurate sampling frame created ☐
- Sample selection process is based on a scientific method and as per the design ☐

6. Keep records of probability of selection at each stage to calculate sampling weights. If a listing exercise is employed for preparing a sampling frame, keep a record of all details regarding the frame, including selection and size of segments, if any.

7. Proper care needs to be taken to ensure that there is no deviation from the proposed design in any form. Field monitoring of listing exercise, centralised system to select segments and sampling units, and use of geo-referenced location data may help achieve this.

8. Consult a statistician to decide and develop an appropriate sampling strategy.

Technology Tips:

- Digital devices can help improve the sampling frame boundary and geo-referenced location. Field teams may save time in reaching PSUs by using PSU geolocation. Online listing (using Google spreadsheet/other alternative tools) of targeted respondents can minimise the listing error and reduce time gap between listing and sampling
- Select PSUs using Google maps
- Use statistical software such as Stata, SAS, R, online sample size calculators or readily available Excel templates for sample size calculation

Machine Learning Tip:

- Use convolutional neural network for identification of structures at the time of sample designing

2.3 Survey Tools

Survey tools consist of both manuals and questionnaires, including a checklist for survey monitoring and reporting.

To ensure data quality, it is important to consider the following points while developing manuals:

- **Develop survey manuals for various levels of field monitoring and survey implementation (for example, supervisor, interviewer). Prepare a survey manual that includes a description of survey procedures, explains the questions and the response recording process. If collecting data on handheld devices, digitise the manuals so that they can be accessed as and when needed. The best practice guidelines for survey manuals are available elsewhere [23-26].**

- **Develop a survey-specific quality assurance manual, which includes instructions on how to do quality control during a field survey.**

A well-designed survey questionnaire should follow the BRUSO model —Brief, Relevant, Unambiguous, Specific and Objective. The quality assurance team shall review the questionnaire's content, formulation, type, sequencing and length of questions to ensure it follows the standard and recommended procedures. The details on standard and common practices followed for developing questionnaires are available elsewhere [27].

Some key points to keep in mind while developing survey questionnaires are:

- **Develop a tool appropriate for the mode of data collection and the type of respondents.**
- **Pre-test questionnaires before implementing in the field.**
- **Translate and back-translate questionnaires to ensure consistency in how the questions are worded.**
- **Include instructions for administering questions where the interviewer may require clarity.**
- **Examine and remove redundant questions, if any.**
- **Use only validated question items/scales.**
- **In a structured questionnaire, review the response codes to ensure the categories are exclusive. Also, ensure the response categories are explained in the survey manual. Use appropriate filters and skips to avoid asking inapplicable questions to respondents and minimise errors during data collection.**

Checklist for survey manual

- Separate manuals available for supervisors and interviewers ☐
- Manuals include detailed instructions and standard operating procedures ☐
- Interviewers' manual explains each question and how it should be administered ☐
- Supervisors' manual contains information on the process of monitoring and feedback mechanisms ☐

Checklist for quality assurance of survey manuals

- Manual includes DQA checklists for pre, during and post data collection ☐
- Instructions on implementing systematic quality assurance procedures ☐
- Includes the roles and responsibilities of quality monitors ☐
- Includes instructions on quality checks to be undertaken during data collection by data quality monitors and explicit actions to be taken in actual setting to ensure data quality ☐
- Templates to enter observations and instructions on how the tool should be administered ☐

Checklist for quality assurance of survey tool

- Questions are according to the objectives of the survey ☐
- Questions are sequential ☐
- Questions are unambiguous ☐
- Follow 'skip pattern' wherever applicable ☐
- Questionnaire is translated and back-translated ☐
- Standard codes are used for response categories ☐
- Data manager involved in the questionnaire review ☐
- Questionnaire is pre-tested and modified based on feedback ☐
  - Feedback from pre-test documented ☐
  - Revisions based on pre-test incorporated ☐
- Include questions that can be used to check internal consistency ☐

2.4 Quality Considerations in Designing Data Entry Applications

In the current times, most of the surveys are using handheld devices such as mobile phones and tablets to collect data from the field. It is important to note that collecting quality data does not only depend on a good survey tool but also on how well the data entry application has been designed and developed. Some key quality considerations while designing and developing a data entry package are:

- **Ensure that the data entry application allows smooth movement across fields.**
- **Ensure both soft and hard checks as well as skips and filters are implemented. Automated skips between variables should be programmed.**
- **Define possible ranges/response codes for all data fields.**
- **Include instructions for each question as defined in the hardcopy of the questionnaire and survey manual.**
- **Include customised warnings/error messages to prevent any inconsistencies.**
- **Incorporate the survey manual within the data entry package that is easily accessible to the interviewer.**
- **Use colour coded instructions to draw the interviewer's attention.**
- **For responses which can be represented pictorially, include relevant images alongside the response categories.**
- **If some of the questions are related and consecutive, group them in one screen.**
- **Collect paradata information such as keystrokes, timestamps for each question and audio recordings (on a random basis).**

- Monitor the response time for each question. If a respondent answers a question faster than the expected response time, include warning messages for the interviewer to improve questioning.
- Ensure that the supervisor has the necessary access for review of data before it is uploaded on the server.
- Collect GPS data at the start and the end of the interview.
- Ensure that both paradata and raw data are linked to the quality monitoring dashboard for real-time monitoring.

Technology Tips:

- Some of the freely available software for developing data entry applications are: CSPro, SurveyCTO, Epi-Info, KoBo Toolbox, ODK and ONA
- These software can also be used to build data quality assurance applications
- Use dedicated servers for data security

2.5 Quality Considerations When Recruiting Survey Investigators

The selection of survey investigators is an important step in obtaining high-quality data. Highly motivated, well-trained field workers are essential for a successful survey. Survey investigators should work in a team comprised of a team supervisor, data quality observers/field editors and a number of male/female investigators as per the need. In surveys that involve anthropometry and biological sample collection, a measurer, an assistant to the measurer and a biomarker specialist should also be included. For quality assurance, the key points to consider when recruiting data collectors are:

1. **Qualified male and female candidates for field staff positions for gender-matched interviews**
2. **A standard screening process for assessing candidates for recruitment**
3. **Fluency in writing and speaking the language in which the interviews will be conducted**
4. **Assess the individual's numerical ability by conducting a short written test on simple arithmetic**
5. **Prior experience in conducting similar surveys can be a factor in determining their suitability for supervisory positions**
6. **Additional staff to account for staff turnover post training assessment and during fieldwork**

Checklist
for field staff recruitment

- Appropriate educational qualifications ☐
- Fluency in local language ☐
- Recruit 10-20% more staff than required in the field ☐
- Recruit gender-matched investigators based on study objectives ☐

Technology Tip:

- Use online platforms for submission and screening of applications

2.6 Training

Training of the field staff is an important part of the survey to familiarise them with the SOPs before they start data collection. In case of large-scale surveys, it may not be possible to train all individuals in one batch. It is recommended that not more than 40 trainees should be trained in one batch. In such a scenario, a cascade training model may be adopted.

The quality assurance for field staff training can be done as follows [28-32]:

1. **Undertake pre and post-training assessments with investigators and supervisors to gauge the level of knowledge on survey tools and processes.**
2. **Conduct mock tests and role plays at the end of each questionnaire section.**
3. **Orient investigators on key terminologies used in the survey and ethical issues in data collection.**
4. **Trainees should be asked to demonstrate their learnings through group discussions on key topics, mock interviews and role plays.**

5. Observe difficulty in the reading of questions and make necessary changes to the questionnaire.
6. Observe whether the investigators have trouble in understanding the questions. Reorient investigators, if required.
7. If possible, have an independent observer to assess the training quality at the end of the day's training.
8. Conduct daily debriefing session with investigators and supervisors to assess the gaps in training delivery.

Checklist

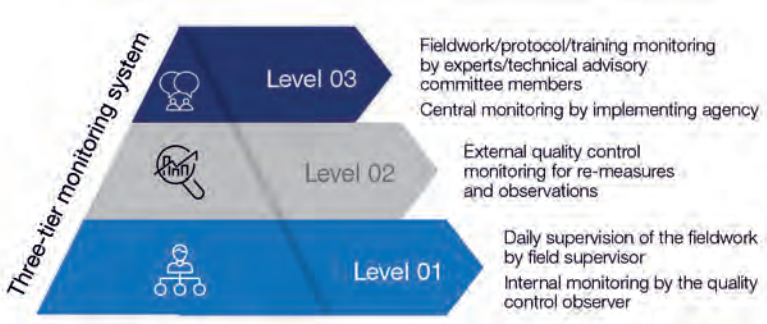
- Trainees attended all sessions on all days ☐
- Training covers all topics/sections, including demo sessions ☐
- Trainer trained in TOT or is part of the core project team ☐
- Training agenda followed completely in terms of topics and time ☐
- Lectures on ethics, sexual harassment and sensitive subjects organised ☐
- Training includes mock interviews, role plays and adequate practice sessions ☐
- Pre and post-assessment conducted with participants ☐
- Field practice and feedback sessions conducted ☐
- Pilot testing of the procedures for biological sample collection, storage, transportation and analysis during training ☐

Technology Tips:

- Use recorded videos to standardise training
- Use recordings of the training sessions to review and assess the quality of training

2.7 Preparatory Steps for Quality Assurance of Anthropometric Measurements

The most common anthropometric data measured to determine the nutritional status of an individual are weight, height, subscapular and triceps skinfold thickness (SSFT and TSFT, respectively), and waist and mid upper arm circumferences (WC and MUAC, respectively). Common errors in measuring these involve body positioning, locating and marking the body landmarks. Errors are also commonly made in reading and recording measurement results.



Checklist

- Monitoring and reporting systems have been defined ☐
- Selected qualified personnel are trained and standardised to take anthropometric measurements ☐
- Job aids and manuals (including videos) are prepared ☐
- Manuals have detailed instructions for measurement, equipment calibration, care and maintenance ☐
- Standardised equipment procured for the survey ☐
- Equipment calibrated as per protocol ☐

Technology Tip:

- A standardisation software package for anthropometry can be used for calculating the technical error of measurement scores during training

2.8 Preparatory Steps for Quality Assurance of Biological Sample Collection

Biomarkers are biochemical, functional or clinical indices of an individual's health status. They are required to support evidence-based clinical guidance for health programmes and policies to improve the health status of a population. A biomarker can be any biological specimen that is an indicator of the health status and can be estimated from various appropriate biological materials like blood, urine, feces, tissue, saliva and hair.

Key steps to be taken prior to initiating data collection for ensuring data quality in case of biomarker measures are:

1. **Laboratories that have internal and external quality control procedures, have ability to meet sample collection, transportation and processing of sample as per protocol specifications, accredited to guarantee uniform sample processing, and use state-of-the-art technology for data management and reporting should be selected.**

2. **Recruit and train phlebotomists who have at least a Diploma in Medical Laboratory Technology. Experience in sample collection in similar surveys should be a desirable qualification.**

3. **In large-scale surveys, in which instant biomarker tests are done in the field (for example, finger-stick blood collected for anaemia testing, malaria testing, blood glucose), a phlebotomist may not be needed and a trained health investigator with suitable background and experience may be sufficient.**

4. **In case multiple laboratories are involved, clearly define equipment and consumables to ensure standard and uniform materials are used across different survey locations. If resources permit, it is recommended to opt for central procurement. Ensure regular calibration of equipment/tools.**

5. Use standard, internationally recognised analytical methodologies in the laboratory for biochemical analysis. Please see Biomarkers of Nutrition for Development (BOND) [33], European Registration of Cancer Care (EURECCA) [34] and National Health and Nutrition Examination Survey (NHANES) [35] for appropriate methodologies.

6. Prepare SOPs with detailed instructions to achieve uniformity in sample collection (time of collection, materials needed for collection, devices and prerequisites like the fasting status of the respondent), storage (storage temperatures, sorting conditions), transportation (temperature requirement during transportation till analysis) and processing including the time within which the samples should be processed.

7. Ensure that the data collection teams are well-acquainted with the standard operating procedures.

8. Engage quality control laboratories for comparison testing where 5% of all samples can be randomly selected and sent for consistency checks and quality assurance.

9. Examine if pilot testing of the procedures for biomarker sample collection, storage, transportation, processing and analysis has been carried out and the observations have informed the updating of standard operating procedures.

10. Engagement of an external monitor/expert to guide the quality control of laboratories can be helpful.

11. Check that appropriate database templates/formats and information systems are available to capture information at all stages starting from the pre-analytical phase of sample collection to the laboratory analysis and reporting.

Checklist

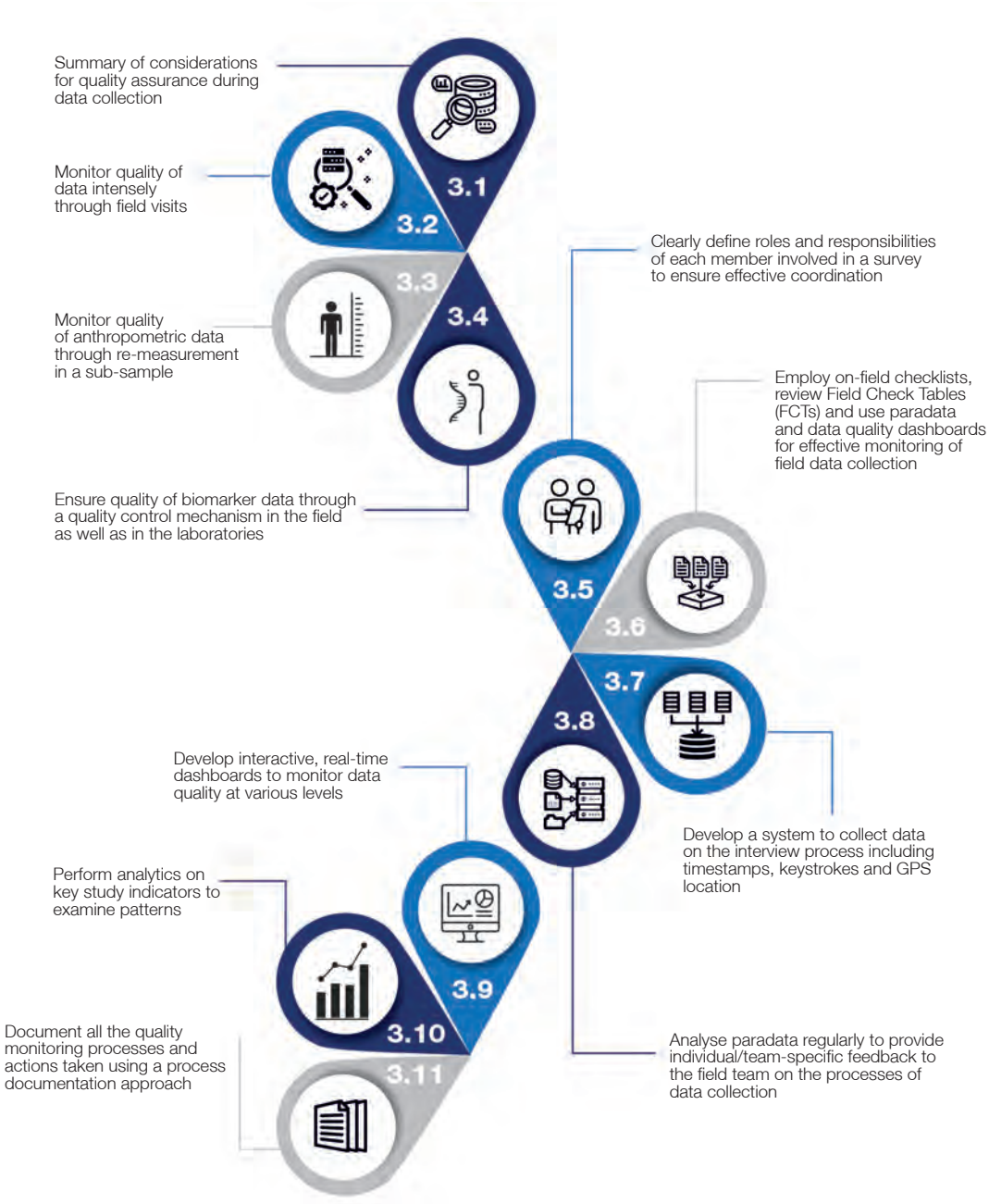
- Laboratories selected have internal and external quality control procedures ☐
- Phlebotomists qualified with a Diploma in Medical Laboratory Technology recruited ☐
- Standard and uniform materials are procured for use across survey locations ☐
- Standard internationally recognised analytical methodologies are used in the laboratory for biochemical analysis ☐
- SOPs are prepared with detailed instructions for sample collection, storage and transportation ☐
- If possible, employ quality control laboratories for comparison testing ☐
- Appropriate database templates/ formats and information systems established for data capture ☐

This page has been intentionally left blank

3. Implementation of Quality Assurance Activities During the Data Collection Phase

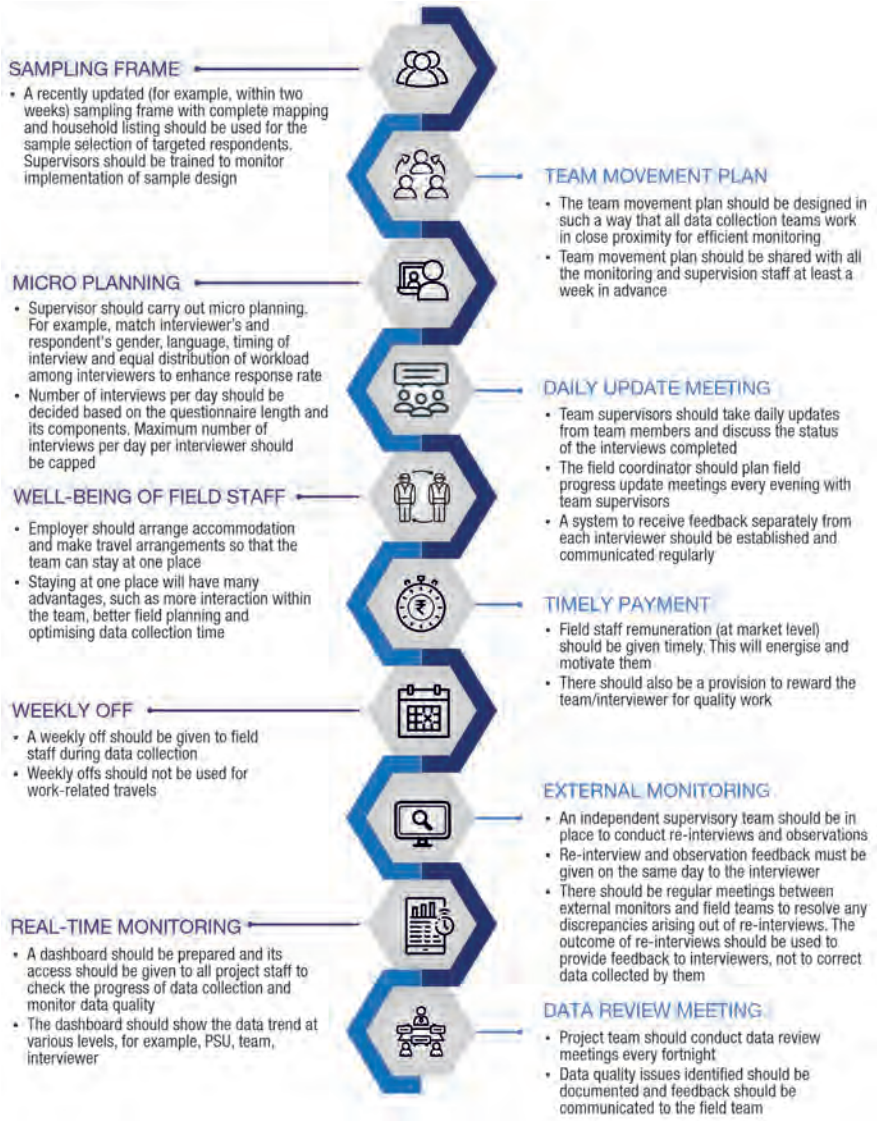
Quality control during data collection is the most important part in a survey. This section of the document outlines 'what' quality assurance parameters should be considered during the data collection phase of the survey and 'how' they can be implemented. It lists out the quality assurance steps, tools to monitor survey, anthropometric and biological data collection, and provides guidance on the coordination mechanism between the quality assurance team and the survey team. It also describes how to use paradata and data quality dashboards to monitor data quality during field survey.

Below are some of the key principles that can be followed during the data collection phase:



3.1 Summary of Considerations for Quality Assurance During Data Collection

Data quality can be enhanced by considering the following points during data collection and should be a part of field operation strategies. The field operation staff (field coordinator and team supervisor) should strictly monitor the following aspects during data collection.



3.2 Steps for Monitoring Survey Data Collection Quality

Monitoring data collection in the field is necessary for improving its quality. With the technological improvements in surveys for data collection, there are online and offline checks recommended to further improve data quality. Following are the recommended steps for monitoring survey data collection:

- **Each team member in the survey has a brief outline (one page or less) highlighting responsibilities they have in the survey.**
- **Each team member is provided with briefs that contain standard definitions, reference interview tool documents and manuals to guide during data collection.**
- **Observe the approach, selection of respondents, consent taken, privacy maintained and sensitivity of the interviews during data collection.**
- **Observe the interviewers regularly during the data collection period, more frequently at the start and towards the end of the survey.**
- **Review the completed interview while the survey team is still in the vicinity of the PSU.**
- **Re-interviews and observations should be carried out randomly during data collection, and feedback must be given to team members on the same day.**
- **Observe data upload and filling of field checklists.**
- **Field check tables generated by the online system (or from the central data quality assurance team) should be regularly reviewed and discussed with the field staff.**

Daily data quality checks

Whether the data is captured electronically or in a paper-based format, data quality should be assessed daily for:

- Missing or duplicate data
- Consistency in the information that is subject to desirability bias – for example, age, reporting of substance use amongst many others
- Consistency in responses between the related questions, for example, age of the woman and number of children, income and expenditure

Checklist for team supervisor

- PSU progress sheet maintained (investigator-wise sample assigned and sample collected) ☐
- Track record maintained for incomplete interviews and planning for re-visits ☐
- Re-interviews and observations conducted for each team member and feedback provided ☐
- Daily progress report sent to the field coordinator and plans for the next day are discussed ☐
- Data is reviewed daily before being uploaded on the server ☐
- Mop-up, if required, is done before leaving the field ☐

Checklist for field coordinator

- Team movement plan in place ☐
- Logistics are arranged for teams in advance ☐
- Coordination and reporting mechanisms established ☐
- Daily briefing with the teams based on the FCTs ☐
- Documentation of quality processes maintained ☐
- Frequent unannounced monitoring visits/reviews conducted ☐

Checklist for data collector

- Record of each assigned interview maintained in the assignment sheet ☐
- Survey protocols are followed ☐
- If applicable, calibration of anthropometric equipment carried out as per protocols ☐
- If applicable, biological samples collected as per protocols ☐
- Data collected is reviewed daily ☐
- Data uploaded on the server daily, post review by the team supervisor ☐

Technology Tips:

- A DQA monitoring application can be developed to compare original data with the re-interview data and list out discrepancies, if any
- Discrepancies are displayed in a report format for each investigator
- SMART Innovations (https://smartmethodology.org/) has the features to check data quality on a real-time basis

3.3 Steps for Monitoring of Anthropometric Data Quality

Adequate and regular monitoring during data collection is critical to prevent errors while measuring and recording anthropometric data. Data collection should be monitored in the field as well as at the central office. Following are the recommended steps for monitoring anthropometric data collection in surveys:

- **Ensure the correct age of the respondent is recorded, particularly for children.**
- **Ensure a balanced workload is assigned to each team so that a reasonable number of households can be visited by them without facing fatigue caused by excessive workload.**
- **To prevent measurement errors, ensure that the anthropometry equipment is set up in each selected household as per the standard protocol before initiating measurement (the detailed procedure for setting up equipment is available elsewhere) [36,37].**
- **Ensure that the equipment is calibrated before starting data collection in the field and regularly thereafter as per a specific schedule, depending on the equipment. Ensure that a calibration log is maintained by the anthropometry team for review.**
- **Ensure that the equipment care and maintenance protocols are followed throughout the survey.**
- **While taking anthropometric measurements, ensure that all objects from hands and wrists of the respondents, such as watches, bracelets, chunky rings, are removed as these might affect precision of measurement.**
- **If there is more than one respondent from the same household, ensure that the measurements are taken one at a time to avoid errors that can be caused by mix-up in data recording.**
- **Avoid parallax error. Ensure the measurer reads the measurement with his or her line of sight directly in front of the value rather than at an angle or from even slightly off the side.**

- Re-measurement to assess accuracy: (1) Blinded re-measurement is recommended on randomly selected sub-samples that have already been measured as part of the survey sample. (2) Flagged re-measurement is recommended for flagged data/implausible values. Re-measurement should be done using the same type of calibrated equipment and standard measurement methods used for the initial measurement.

Checklist

- Age of the respondent verified ☐
- Set up of anthropometric equipment for measurement as per the protocols followed ☐
- Routine calibration of anthropometric equipment done as per the schedule and calibration log maintained ☐
- Job aids and manuals are available ☐
- Parallax error avoided ☐
- Re-measurement done to assess accuracy (blinded random re-measurement of sub-sample and/or flagged re-measurement) using the same type of calibrated equipment and standard measurement methods ☐
- Measurement errors detected within acceptable limits ☐

Technology Tip:

- A software package can be developed to enter re-measured data while in the field and compare between original values entered by the field team and re-measurement data entered by the quality control team

3.4 Steps for Monitoring Biomarker Sample Collection, Storage and Transportation Processes

Collection of samples for biomarker analyses in field surveys and obtaining reliable results is challenging as samples have to be collected and transported from the field to the laboratory under various conditions. Although in many large-scale surveys, biomarker tests are done in the field itself, if the survey includes biomarker indicators, a rigorous quality assurance procedure needs to be established using standard internal and external quality assurance procedures. Some important steps are as below:

Depending on the biomarkers proposed to be tested in the survey, develop a standard operating procedure on the sample volume and types needed (plasma/serum; trace element free), processing methods, aliquots and storage conditions, and maintenance of relevant time record for each of these variables.

- **Ensure selection of eligible children/adults by checking if the phlebotomists have identified the correct households and eligible respondents as per the sampling list. Ensure that biological samples are collected from the selected individuals.**
- **Ensure the use of standard equipment and consumables during sample collection to maintain uniformity and standardisation among different phlebotomists.**
- **Check if appropriate instructions are being given to the respondents.**
- **A monitoring checklist including a core set of essential tasks which should be performed during the collection of biological samples, their storage, transportation, and processing can be used to aid in measuring the phlebotomist and laboratory performance.**

- Several biomarkers are adversely affected when the blood samples are stored for a long time period before plasma/serum collection, due to partial/complete cell lysis. Therefore, a careful sample processing plan (often specific to the biomarker of interest) needs to be thought out before the sample collection.

- Keeping the sample in 2-8^{o} C is necessary if longer processing times are anticipated (>6 hours). Therefore, ensure that the cool boxes used for sample transportation maintain adequate temperature for at least 12-16 hours. It is recommended to do a pilot test of the performance of cool boxes/bags by simulating field situations. Ensure that the tubes/containers with samples are placed in cool boxes without direct contact with the ice packs as that may affect sample integrity.

- It is essential to ensure that the haemolysed samples are either excluded or records of those samples are made, for later analysis as needed.

- A separate record of sample collection time, storage temperature, and processing time needs to be maintained to analyse the effects of these variables on biomarker estimates, if any.

- Aliquoting and sample storage: Since repeated freeze/thawing cycles affect biomarker estimates, a careful planning and maintenance of aliquots are often necessary. This will also facilitate storing the samples at the required temperature. Record the date of sample analysis and any repetitive freeze/thaw cycles.

- Check labelling of samples. Appropriate labelling of each sample and aliquot is required for the identification and tracking of biological samples. It is recommended to code the IDs with cryogenic barcodes in large-scale surveys to minimise reading mistakes and simplify sample tracking with the help of scanners.

- A periodic evaluation of phlebotomist's performance is recommended to ensure testing procedures are performed consistently and accurately, with retraining, as needed, based on the results of the assessment.

- **In case of spot testing of samples, such as random blood glucose testing or haemoglobin testing, check if the results are recorded accurately.**

- **Process control:**

  - If samples are processed centrally, it is recommended that they be shipped daily to the laboratory to ensure they reach within 24 hours and the integrity of the samples is maintained. In this case, maintaining an accurate time record of sample collection, sample processing (serum or plasma collection) and biomarker analysis is a must.
  - If several local laboratories are engaged for processing the samples, there is a higher need for quality assurance to ensure that the standard and uniform methodology is used across the laboratory chain.

Choice of methods for biomarker analysis

Ensure the use of most recent and internationally-accepted analytical methods, as it will add value to the data. Also, prepare relevant SOP records for analysis and record the amendments from time-to-time.

- **Internal quality assurance:**

  - Check that the laboratory runs a three-level *in-house* quality control sample after a batch of survey samples.
  - Precision (with the limit of detection) should be measured for all assays at various reference ranges. Attempts should be made to keep the analytical coefficient of variation (CV) for an assay within 5-10%. However, this will vary depending on the biomarker being measured.
  - Wherever available, use standard reference serum samples for biomarker analysis.
  - Every week, a percentage of samples should be split and reanalysed and results should be compared.
  - Blind quality control samples can be inserted into some batches along with the respondents' samples to monitor the laboratory's performance during and after the assays have been completed.

- **External quality assurance:**
  - A subset of samples (at least 5%) can be sent to quality control laboratories for comparison testing as frequently as possible.
  - Recommend the laboratories to participate in the BIO-RAD and US Centre for Disease Control external quality assurance schemes.
- **At the laboratory:**
  - Check if the procedures being used for sample processing are in line with the SOPs prepared for the study.
  - Review laboratory registers to check whether the samples collected match those that have been sent to the QC laboratories.
  - Check the calibration log in the laboratory to ensure devices were calibrated as per the SOPs.
  - Review the results of the comparison analysis carried out between the results from the laboratory and the QC laboratories.
  - Review the data logger (if used) results every week and give feedback to the laboratory personnel. The laboratory personnel, in turn, should provide feedback to the phlebotomists.
  - Check if the internal quality assurance systems are being implemented.
- **Perform data quality checks. Every fortnight, the following data quality checks should be carried out and feedback should be given to the laboratory:**
  - Percentage of implausible values for each biomarker analysed
  - Percentage of outliers for each biomarker analysed
  - Percentage of missing values
  - Percentage of cases where the sample was not sufficient, test not performed or invalid results reported
- **A proper protocol should be written for the disposal of waste. The vacutainer needles should be collected in a puncture-proof container with lid. These should be discarded at the laboratory or collection centre. Wastes like gloves should be collected in a bag with a biohazard symbol. Wastes such as cotton swabs can be disposed on-site in a bin.**

Technology-based monitoring during data collection

- **Time and temperature monitoring with a data logger** is recommended during storage and transportation of samples. Ensure the data logger is placed in the cool bag from the time the first sample is collected in the PSU and switched off just before the sample is removed from the cool boxes for analysis in the laboratory. Thus, the temperature during the entire journey of the sample can be assessed. The data from the data logger should be reviewed, invalid samples should be removed from the data and ongoing feedback should be provided to the laboratory so that such delays are not repeated, and sample integrity is maintained.

- **Text message-based field monitoring:** A text message-based monitoring/alert system is recommended to monitor the daily dispatch of biological samples from the PSU, samples being received at the collection centre and at the reference laboratories where the samples are being analysed.

    - **Step 1:** First message sent by the phlebotomist when the first sample is collected from the first respondent in the PSU.
    - **Step 2:** Second message sent when samples reach the collection centre.
    - **Step 3:** Third message sent when samples reach the reference laboratory where they are processed.

  Whenever there is any breach of time, automatic alerts should be sent to phlebotomists, laboratory staff and quality control monitors so that the reasons for the breach can be assessed and rectified.

- **Computer-based daily sample collection monitoring system** is recommended to assess a mismatch between the number of samples collected in the field and the number reported by the laboratory.

Checklist

- SOPs with detailed instructions are prepared for sample collection, storage and transportation ☐
- Laboratory runs internal quality checks ☐
- A subset of samples is sent to QC laboratories for comparison testing ☐
- Samples collected from selected eligible respondents ☐
- Standard equipment and consumables used for sample collection ☐
- Cool boxes are used for sample transportation to maintain adequate temperatures for at least 12-16 hours ☐
- Appropriate instructions are given to the respondents ☐
- Job aids available with the phlebotomists ☐
- Each sample and aliquot are appropriately labelled ☐
- Results are recorded correctly in case of spot testing of samples ☐
- Correct procedures are followed for sample processing ☐
- Calibration log is maintained, and devices are calibrated as per the SOPs ☐
- Comparison analysis carried out between results from the laboratory and from the QC laboratories ☐
- Time and temperature monitoring are undertaken using a data logger and feedback is given based on results ☐

Technology Tips:

- SMS-based alert system (for example, RapidPro) can be used to track the journey of biological samples from the point of collection in the field to the laboratory where they are analysed. This system should include automatic alerts sent whenever there is breach of time
- A data logger can be used to monitor time and temperature during the sample's journey from field to laboratory. Automated programmes can be developed to analyse data logger data to maintain a log of any breach in temperature and provide immediate feedback
- Cool bags containing biological samples can be geotagged

3.5 Mechanism for Coordination Within and Between the Quality Assurance Team and the Main Survey Team

The DQA teams and main survey teams must maintain a clear understanding of their roles and responsibilities during data collection. The following points should be considered to have the teams work in tandem to ensure data quality:

- **During the investigators' training, there should be a dedicated briefing on roles and responsibilities of the external monitoring team at the time of data collection.**
- **Team supervisors or field coordinators should send their field movement plan, ideally a week ahead to the DQA team so that they can plan their field movements accordingly.**
- **The DQA team should meet the team supervisors regularly to discuss the field team's performance and quality of data collection. They should help the survey teams resolve any problems related to finding assigned households, understanding the questions, recording of responses, dealing with difficult respondents and other issues.**

Technology Tip:

- A field team performance dashboard can be created to monitor the efficiency of the team, workload, field movement, attendance of field staff and weekly-off status. A Google dashboard may be used

3.6 Tools to Monitor Quality of Field Data Collection

There are several tools that can be designed for monitoring data quality depending on the type of survey, the issues that are examined and the levels at which each survey is conducted. Following are the recommended tools for data quality measurement:

1. **Field Monitoring Checklists (FMCs):** FMCs are important monitoring tools that can be utilised by field supervisors or coordinators. FMCs generally consist of information on the number of interviews assigned and completed per day by investigators, non-response rate, number of attempts made to complete the interview and whether consent processes are followed. FMCs can be prepared at different levels to ensure accountability for data quality.

2. **Re-interview tools:** Re-interview tools in the form of back-checks are abridged versions of the main tools and mostly include factual questions that are not time-sensitive or perception-based. These tools are important from the perspective of monitoring to ensure data quality.

3. **Daily debriefing checklists:** Daily debriefing checklists are important to document field level challenges. Further, these checklists aid in addressing the need to retrain or resolve any other field level issues discussed during debriefing sessions. Supervisors should conduct a daily debriefing meeting with data collectors at the end of the day.

4. **Daily SMS updates:** Daily SMS could be another field monitoring tool to ensure quality data collection. It can help team supervisors to complete the PSU progress sheet and provide discussion points for any field level challenge that requires immediate attention and mitigation.

5. **Field Check Tables (FCTs):** FCTs are created considering the key indicators pertaining to study objectives and can be developed for each team and interviewer. FCTs play an instrumental role in data quality assurance. FCTs help monitor the response rate, negative screening, estimates of critical indicators, investigator efficiency and bias. FCT data should be reviewed and discussed with all levels of field staff for better understanding and to provide feedback.

6. **Analysis of key indicators:** In addition to generating FCTs, it is important to undertake additional analysis of key indicators and measures to understand quality of data. Such analyses include reviewing frequency distribution and assessing reliability (test-retest, alternate form and internal consistency) of data. For anthropometry, the standard deviations of the various measures in standardised scales (Z-scores of WHO) are important indicators to assess the quality of data.

7. **Paradata and dashboard:** For monitoring of data and quality assurance, regular feedback from the central office is required to guide the field team. Paradata and dashboards provide the avenue to monitor data in real-time at the central office. In general, process indicators related to efficiency and productivity of survey investigators and other key parameters which define the quality of data are monitored using paradata.

3.7 Use of Paradata to Improve Data Quality

The organisation involved in large-scale data collection may use paradata to improve the quality of data. The term paradata refers to the auxiliary data collected in a survey that describes the data collection process. Paradata can be utilised in many ways to improve the overall data management, leading to better data quality. The usage includes:

1. **Improved understanding of the entire data collection process and ability to assess new methodologies related to data collection**
2. **Real-time monitoring of process indicators during data collection**
3. **Minimising survey errors**
4. **Estimation of overall survey errors post data collection**
5. **Quality control metrics to set the benchmark**
6. **To make evidence-based decisions regarding the survey cost and output achieved**

Process indicators for paradata

Process indicators used in real-time monitoring can enable data producers to investigate survey errors in many different scenarios. The entire set of process indicators generated as paradata can be classified into:

1. **Interviewer data collection status/status of sample**
2. **Interviewer productivity/completed cases**
3. **Dataset representativeness/response rate/non-response/ doorstep refusal**
4. **Tracking negative screening by investigators**
5. **Tracking team movement**
6. **Identification of outliers for key outcome indicators**

7. **Enabling the identification of crisis on the field and implementing evidence-based rapid response**
8. **Comparing the data quality metric to ensure quality output**
9. **Error-cost trade off report that includes survey performance reporting indicators such as time per unit, cost per unit and completion rate at various levels of aggregation**

Type of paradata and their utility

The above-mentioned indicators are broadly classified under five types of paradata.

Type of Paradata	Usage
Audio recordings	Recordings of interactions between the investigator and the respondent (subject to consent) can help understand the reasons for non-response and assess the quality of the interview.
Keystrokes	Keystrokes can help assess whether the recorded answer was changed, which can be a potential measurement error.
Interviewer traits	Interviewer-specific indicators help understand the protocol and process followed. The indicators generated under this type of paradata can help assess negative screening rate and any other anomaly caused by the investigators.
Timestamps	These can help monitor interview time.
GPS data	Geospatial analysis of team movement can help in tracking team/individual movement. This information can help minimise coverage error or bias. GPS data can also help modify the field movement plan to make the survey more efficient over time.

3.8 Type of Analytics on Paradata to Present Data Quality Metrics

Broadly, paradata can be utilised in two ways:

1. **By providing paradata on key elements related to the survey process, the quality of data can be ensured by minimising survey errors during the data collection phase.**

2. **Post data collection, paradata is used to calculate the total survey errors by estimating coverage errors, non-response errors and measurement errors.**

Based on the type of paradata, the following analytics can be used:

1. **Timestamp:** Timestamp paradata can be analysed to understand the total survey duration and to further investigate the time spent by field investigators on a particular question or section. The mean time duration per team/investigator should be a parameter to detect any anomaly in the data collection process.

2. **Negative screening rate:** Negative screening rate is calculated as the total number of sections in a tool skipped by an investigator compared to the total number of sections in the tool. This will help understand any behavioural issues or the need to re-orientate the staff.

3. **GPS data:** The Global Positioning System data can be used to track team movement, increase efficiency of survey implementation strategy, coverage bias and any clustering issue in the estimates.

4. **Keystroke:** Keystroke data are useful to monitor the changes in the response punched in by the field investigator. Keystroke data can be analysed by taking the frequency to understand where the maximum punches for a particular variable took place. Higher frequency of keystroke data can also indicate that the investigator/respondent might have found it difficult to understand the question.

5. **Audio trails:** Audio trails can be useful to understand the data collection process, monitor the ethical aspects and assess the questions that are asked in the prescribed manner.

The results from the analysis of paradata may be documented in the data quality reports as a part of the survey documentation.

Technology Tip:

- Use software such as Stata, Excel, SPSS, Python, R, CSPro for analysis of paradata

3.9 Using Dashboards to Monitor Data Quality

In a large survey, it is important to monitor data frequently (or) on a real-time basis. Real-time monitoring reduces the human effort in analysing and monitoring the long line of communication that can hamper timely feedback to the field teams. It may be monotonous and cumbersome to monitor tabular data regularly. Therefore, survey monitoring should be done using real-time data quality dashboards. A data quality dashboard is an information management tool that continuously monitors data quality visually, enables tracking trends and allows quick analysis. The dashboard should have provisions to visualise these indicators by investigator, survey team, period of survey and area of survey. Having a dashboard for data quality monitoring also helps in a continuous feedback process for enumerators and improves their accountability and motivates them to be careful about the data collection process. A dashboard should be visually appealing and contain the following information:

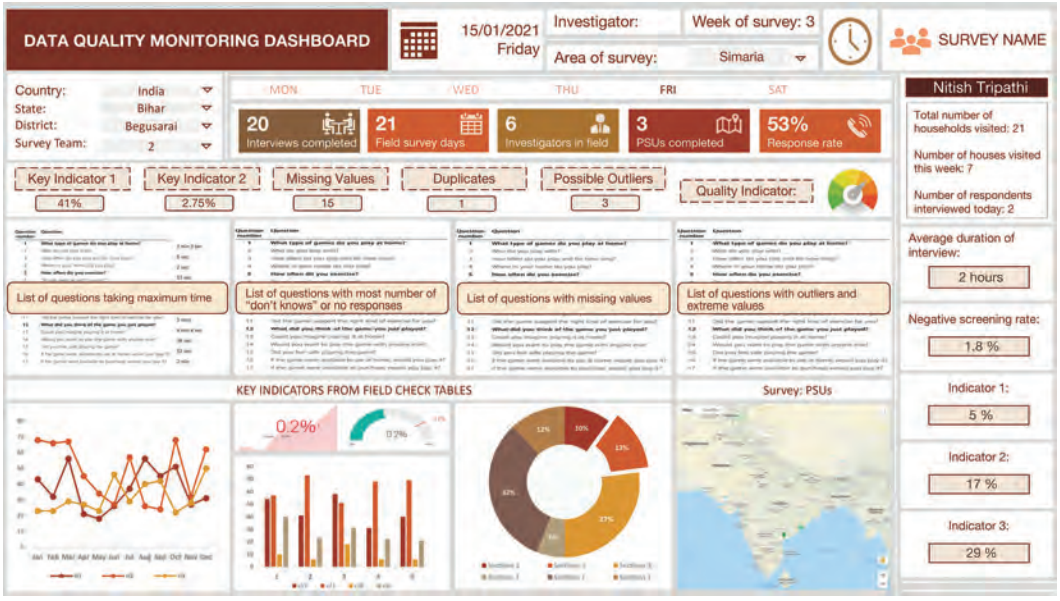
- **Day-wise interview status including response rate**
- **Interviewer productivity**
- **Average duration of the interviews**
- **Negative screening rate**
- **Key indicators from the field check tables**
- **Indicators which have a high risk of error during data collection**
- **Outliers and extreme values in key survey indicators**
- **Number of missing values in key indicators**
- **Observations which are duplicates or likely duplicates**
- **List of questions taking most of the interview time**
- **List of questions with most 'don't know' and/or 'no answer' responses**

A data quality dashboard should have options to demonstrate the time trend of the above quality monitoring indicators and should also enable filtering at various levels of data collection. Additionally, it should be prepared in a way that a non-technical user can navigate it easily and be able to clearly understand the data quality issues as and when they creep in during data collection. The data quality dashboard

should be made available to all relevant supervisory positions starting from the field supervisor with viewing controls for different levels. For example, a field supervisor should be able to view only the performance of his/her team members; whereas someone sitting at the central office should be able to view details of all teams.

For a real-time data quality dashboard, it is important to choose the right data collection platform. One should prefer a platform that allows connection between the data storage server and the dashboard creating platform. While there are several dashboard creating platforms such as Power BI, Dashit and Google Dashboard, it is important to choose one which can be smoothly managed and easily handled by the study team.

A sample data quality monitoring dashboard



Technology Tip:

- **Use dashboard creating platforms such as Power BI, Tableau, Dashit, Google Dashboard to see the real-time changes in the data and monitor performance efficiently**

3.10 Indicators to Measure Data Quality for Providing Feedback to Investigators During Data Collection

During data collection, the survey agency/nodal institute must ensure that regular quality check reports are prepared and sent to the field teams. Different indicators should be assessed and feedback on quality aspects must be sent to observers in the field so that the investigators are appropriately debriefed.

In population and health surveys, some key indicators for feedback include [38-40]:

- **Household completion rate: Out of the total number of eligible households; the percentage of households completed, refused, dwelling vacant or destroyed or not found**
- **Number of household members at home with their age-sex composition**
- **Completeness of age and age heaping in all collected age variables – age, age at marriage, age at first child, age at death (if any)**
- **In case of a child, the percentage of date of birth information obtained from birth certificate, vaccination card, caretaker's recall, or other sources**
- **Completeness of length and height in case of anthropometric measurements. Standing/lying position for length/height in case of a child. Digit heaping in height and weight measurement**
- **Average time taken per schedule**
- **Frequency at which equipment is calibrated**
- **Number of interview schedules filled per investigator per day**
- **Missing information and skipping pattern followed and outliers identified**

In case of health and disease-specific surveys, some key indicators for feedback include:

- **Number of patients in a household with their symptoms of health problems**

- Pattern in reported days between onset of symptoms and diagnosis of disease
- Pattern in reported days between diagnosis and treatment seeking
- Treatment seeking (government vs private health facility)
- Number of new and relapse cases of infection diseases

The data quality feedback provided to the investigators must contain the following measurements:

- Review of filled interview schedules for missing information, outliers and skipping pattern
- Re-measurement of sub-samples can be performed while the survey team is in the field
- List of investigators with missing information, outliers and skipped questions

Technology Tip:

- For real-time monitoring of data quality, a dashboard for supervisors and field coordinators can be prepared utilising information on completed cases, non-response, missing values, negative screening rates, interview completion rates, average time taken to complete interviews and so on

Machine Learning Tip:

- Use decision trees, neural networks, SVM, K-means for classifying the number of times 'don't knows' and 'skips' are used by either the interviewer or the respondent

3.11 Documentation of Data Quality Assurance

Documenting the survey process, specifically the quality assurance protocol is an integral part of survey implementation. A document on data quality should include not only procedures that ensure data quality before and during the survey but should also include a detailed description of different quality parameters assessed using appropriate statistical procedures.

A document on data quality helps data users understand the strengths and limitations of data and also enables them to derive appropriate conclusions from the data. At the same time, it also helps other data producers reproduce similar data by implementing similar data quality assurance mechanisms. It further helps carry out a comparison of data across surveys.

Following are the recommended areas for data quality documentation:

- **Data quality assurance practices that were followed before and during survey data collection**
- **Data quality observations from the field**
- **Communication notes on data quality steps and feedback meeting reports**
- **Citations for the measurement approaches used**
- **Key results and interpretation from the analysis of survey paradata**
- **Key results from the analysis of comparison carried out between raw data and field check tables**
- **Data validation and screening procedures implemented during and post survey**
- **Detailed data profiling including the transformations and adjustments made to the raw data**
- **List of variables identified with flagged observations and if possible, detailed documentation on the reason behind such data flags**
- **Merits and shortcomings with specific data points and recommendations for statistical analysis**

Documentation of survey data quality should start from the preparatory phase. It could be designed as a process document where details are noted on each step of survey planning and implementation from the perspective of data quality. Moreover, one person should be exclusively assigned to document all quality aspects by interacting with field teams, survey managers and other research team members.

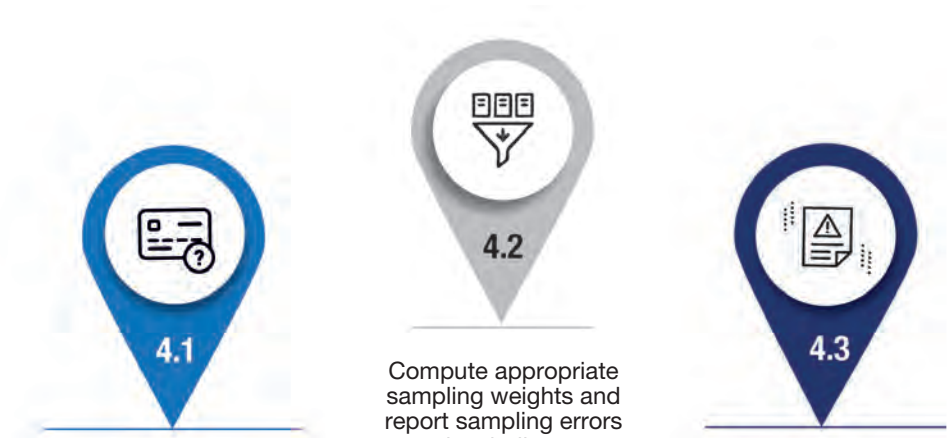
Technology Tip:

- **Use digital platforms (for example, WhatsApp, Slack, Chanty, Flock, Hangout) to share audios/videos, photos/screenshots, text messages, conduct group discussions and also to document DQA reviews and actions taken within the project team**

This page has been intentionally left blank

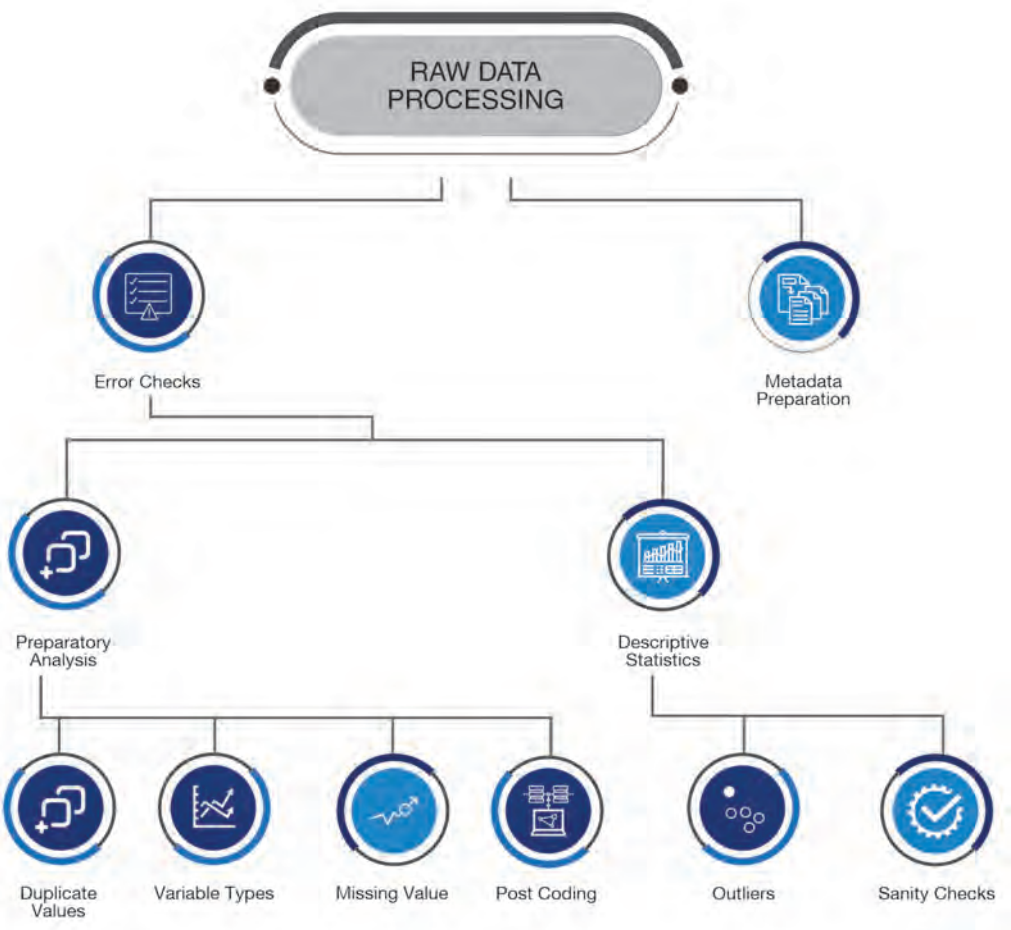
4. Data Quality Assessments Post Data Collection

Once data is collected, it is important to review the data, and undertake analytics to examine data quality parameters. This section of the document outlines the different techniques that can be employed for assessing quality of data, and estimation of sampling and non-sampling errors. It also guides computation of sampling weight and application of different machine learning techniques in surveys. Below are some of the key take-aways from this section of the document:



4.1 Post Survey: Profiling Survey Data

One of the most critical steps in data quality assurance is processing of raw data to make it suitable for analysis. There are two critical aspects in processing raw data: (a) checking errors in raw data and (b) preparing metadata. Even before checking errors in raw data, a key step involves reconciliation of field data. Handling large quantity of data can get tricky and tedious, both for the data analysts as well as



the analytical tools or software in which the data is being analysed. It is, therefore, of utmost importance to perform systematic checks that are mentioned below on snippets of data every time the dataset is updated to avoid any error creeping in. Time and again one must also compare the new version of the file with the older ones and check the total number of rows and columns in each to make sure no data is lost or converted or corrupted in the process.

Checking errors in raw data

It is inevitable to have some errors in raw survey data. Therefore, before it is put to use, a data analyst needs to review the robustness of all data fields. It is the stage where one needs to perform a few checks to ensure that the data collected adheres to good quality. This phase is the longest, starting from data cleaning, i.e., removing duplicates, flagging missing values, and outliers, to exploratory data analysis where one attempts to examine distribution of key variables. This chapter will outline key steps involved in finalising a raw data set.

1. **Preparatory Analysis:** Preparatory analysis is the process of cleaning and transforming raw data prior to processing and analysis. It involves reformatting, correcting and combining the data sets to enrich data. For example, the data preparation process usually includes standardising data formats and enriching
source data.

    - **Duplicate observations:** The first step once data collection is finished is to check whether one or more observations in the data are duplicated, i.e. have the same values across all variables, including values in the unique identity column. If a duplicate observation exists in a data, one should investigate whether an incorrect entry has been made or a genuine duplicate. If it is a genuine duplicate, then it is advisable to drop one of the duplicated cases. Ideally, data should be unique, and no two rows should be same in a dataset. Though one can easily delete rows that are duplicates directly in a data processing software, one can also train machine learning models like fuzzy logic with historic data to predict whether any new data is a duplicate of the historic data or not.

    - **Variable type:** One of the common issues observed in raw data is in the type of variable (data), i.e., data being expected in numeric but

available in string (text) or vice-versa. The data analyst should review all the variables and their data types. If any variable is not as per the expected data type, it should be transformed while finalising the data. Similarly, certain fields can also have a limit to the length of the strings which might not permit characters beyond that thereby leading to incomplete data entry.

- **Missing value:** Check whether the values that are missing from the data are the ones which were specified at the planning phase of the survey itself and were expected to have missing values in them (i.e as per skip patterns in the interview tools). Given that most of data is collected in handheld devices, the chances of missing values are rare these days, However, if still there are missing values, data analyst needs to check with the field team the reason behind missing values. Such reasons should be documented (if possible, in the data itself). In some instances, people prefer imputing missing values, in such cases, it is better to use multivariate imputation method. For such imputation, one should create relevant variables which can help determine the outcome of a missing observation.

- **Post coding:** While most surveys have structured responses, there is always a possibility of responses in other categories. A common error that happens while collecting data is that some of the text entered in "other" field corresponds to one of the pre-coded response categories. Therefore, it is important to post-code all such text values to their respective coded response categories. Also, it may happen that for some variables, there is a heaping of a particular response in "other" category and it is not possible to assign them to a pre-coded response category. In such a case, another code category for that specific other value should be created in the data itself.

2. **Examine descriptive statistics:** The next logical step would be to examine descriptive statistics of all variables in the data. Descriptive analysis helps to identify the data quality issues occurred during the data collection process. It can be elucidated through data visualisations (like graphs, charts), measures of distribution (percentiles, quartiles, quantiles, skewness and kurtosis), measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation, range,

interquartile range and coefficient of variation). This would facilitate understanding the performance of data.

- **Outlier detection:** Examining frequency distribution also helps in identifying any extreme (outlier) value a variable may have in the data. While outliers can be possible, it is important to verify if it is resulting from incorrect entry at the time of data collection. If possible, the data analyst should get in touch with the data collection team to re-confirm the values. In case it is not possible to consult the field team, the data analyst should flag such observations. The traditional ways of understanding whether there are outliers in a dataset are by creating boxplots or by calculating and comparing the values of mean and standard deviations of the residuals. One can also use machine learning techniques such as isolation forest, minimum covariance determinant, local outlier factor, one-class support vector machine to detect outliers.

- **Sanity checks:** While during data collection, several range, skip and simple validation checks are included in the data entry program, it is not possible to include all relational checks during the data collection. Before finalising the data, one should check for any out of range values and internal inconsistencies in the data. The data analyst should verify if the data in a variable is consistent with other related variables in the same data. To check consistency of the data, one should write program in data processing software and document if any inconsistencies are found. If the data analyst decides to change any value, it should be documented and rationale behind the change should be mentioned.

Preparing metadata

Metadata is the documentation of the data including important details about the data, the instruments, protocol information, analysis approach and survey tool details. However, in most cases metadata is not well documented and therefore the potential of having good metadata is often not realised. Metadata should ideally answer all the what, why, where, who, how questions about the survey. Metadata for survey data should include definition of all variables, description of all coding values, date of data creation, information about data custodian, and documentation of specific data issues.

Whenever there is human intervention, the possibility of errors creeps in, therefore, while performing these data checks and preparing metadata, one must be extremely careful that no further distortion to the data is made. Efficient data analysts should perform these checks, reconcile carefully, and document all the changes made to the original raw data for future reference.

Technology Tip:

- Use of open sources such as Python, R, etc. and statistical packages such as STATA, SPSS, SAS, etc. can help in cleaning and summarising the data

Machine Learning Tips:

- Use optical character recognition to scan documents to text without human intervention thereby reducing the chances of introducing errors
- Use functions like isolation forest, extended isolation forest, autoencoder neural networks, replicator neural networks, one class SVM, clustering methodologies for detecting outliers
- Use natural language processing for post coding
- Use random forest for missing value treatment
- Use clustering methodologies for data cleaning

4.2 Sample Weights and Sampling Errors

A desired selection of sampling units is EPSEM sampling, which means such sampling will result in the population elements having equal probabilities of being included in the sample. In other words, in EPSEM sampling, each sample unit is self-weighting or the reciprocal of the probability of selection of each element in the selected sample is the same. However, in practice, adopting an EPSEM design may not be possible and this is fine as one can simply address the problem of unequal probability of selection by applying appropriate sampling weights, discussed below.

In a sampling design, if units are not selected with EPSEM, the sample mean will not be an unbiased estimate of the population mean and it would become imperative to use some weights to take care of the bias in estimation. In probability sampling, the probability of selection of a unit is known and the sample weight for each unit can be reciprocal (inverse) to its selection probability. Multiplying the variate values with their respective weights will provide an unbiased estimate of a parameter. Besides, computing sampling errors of key outcome indicators and reporting them are important steps of documentation of data quality.

Basic principles to compute sample weights and sampling errors are:

1. **Assuming that the sample design adopted is a multi-stage design, the first step is to calculate probability of selecting a unit at each stage. For example, in a two-step design, calculate probabilities of selecting a first stage unit and a second stage unit. An example of a two-stage design is a household survey where a village/ward is selected in the first stage from a sampling frame of all villages/wards in a state or country and then select a household for interview in the second stage. In some studies, a selection may also occur within a household making it a three-stage design.**

2. **The next step is to calculate the overall probability of selection of a study unit (for example, a household, a woman within household) by multiplying all probabilities calculated thus far. The reciprocal of this overall probability is the sample weight.**

3. It is a good practice to incorporate sample response rates into the weight calculation to adjust for any bias due to differential response.

4. Finally, one may choose to normalise weights before applying it to the data for analysis.

5. To calculate sampling errors, add information on sample design including stages of selection and stratification to the dataset.

6. Another important indicator to report is the design effect, which is, defined as the ratio between the standard error using the sample design adopted in the survey and the standard error that would result if a simple random sample had been used.

Checklist

- All required information for calculating weights available with the project team ☐
- Considered all probabilities of selection at each stage ☐
- Compared weighted population with census population on important stratification variables (for example, rural/urban, ST/SC, literacy) ☐
- Normalisation of weights ☐

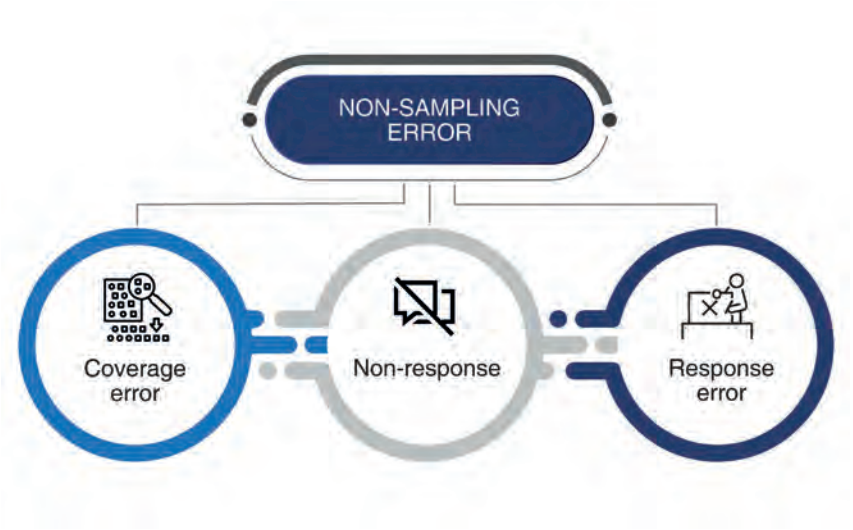
Detailed discussion on sampling weights can be found elsewhere [20, 22, 41].

Technology Tip:

- Sampling errors can be calculated using statistical software (for example, Stata, SPSS, SAS) by specifying the sampling design

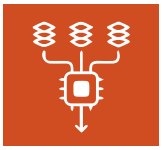
4.3 Data Quality Metrics: Calculation of Non-Sampling Errors

Non-sampling errors are caused at the time of data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions and data entry errors. Non-sampling errors can be classified into three categories [42] :



Non-Sampling Error	Assessment Methods
Coverage error: This is the lack of one-to-one correspondence between the elements in the target population and the elements encompassed by the sample selection procedures used in the study.	Coverage error = under coverage + over enumeration Where under coverage indicates omission of households and over enumeration indicates re-enumeration of sample. Post enumeration quality check is another method for assessing coverage error.

Non-Sampling Error	Assessment Methods
Non-Response Error: Non-response arises when households or other units of observation which have been selected for inclusion in the survey fail to yield all or some of the data that were to be collected. Two major categories on non-response may be identified as non-contact and refusal. Non-contact occurs due to difficulties in accessing sample units, failing to contact respondents, failing to gain cooperation. Refusal occurs when respondent deny providing information.	Non-response error is assessed by the response rate. This is calculated as "the number of eligible sample units who responded to survey divided by the total number of eligible sample units".
Response error: This occurs due to collection of invalid or inappropriate data from sample elements which lead to inconsistency in data, missing values, and outliers.	Response error can be reduced by the length of recall which is the time elapsed between the date of a particular event or transaction that occurred during the reference period and the date on which a respondent is asked to recall it. Missing data in general hampers the reliability of estimates and may be treated as a response error. Methods to impute the missing values can be used to minimise the response error. Missing data can be replaced by imputation using various methods: • Cold deck imputation [43] • Hot deck imputation [44] • Random imputation [45] • Mean value imputation [46]



5. Use of Machine Learning Techniques in Improving Data Quality

In the last decade, there have been significant technological innovations, especially due to the application of Artificial Intelligence (AI) including Machine Learning (ML). Even in survey data collection and management, there is a range of machine learning and artificial intelligence techniques that helps improve data quality in a cost-effective way. For example, in a field survey, identifying the right household to interview was earlier a manual job and often led to confusion and wrong selection of households. With the use of geo-spatial and ML algorithms for image recognition, one can improve the accuracy of household selection as well as save cost of listing each household. Similarly, once data collection is over, one can use isolation forest to find potential anomalies in the data. This chapter will discuss some of the applications of ML algorithms that can help to improve data quality.

PRE-SURVEY:

ML Technique	How it Helps?
Convolutional Neural Network (CNN) [47]	Captures local details and extracts notable image features, which help in identifying the targeted households that the interviewer has to visit

DURING SURVEY:

ML Technique	How it Helps?
Support Vector Machine (SVM) [48]	Classifies times, don't know responses, and skips used by the interviewer and the respondent
Neural networks [48]	Classifies times, don't know responses, and skips used by the interviewer and the respondent
Decision trees [48]	Classifies times, don't know responses, and skips used by the interviewer and the respondent
K-means [48]	Groups interviewers' times, don't know responses, and skips usedby the interviewer and the respondent

POST SURVEY:

ML Technique	How it Helps?
Isolation forest [49]	Isolates anomalies based on random features and values within the range of the data
One Class SVM (Support Vector Machine) [49]	Learns the boundaries of the points and then classifies the points that lie outside the boundaries
Extended isolation forest [50]	Isolates anomalies by selecting a random intercept chosen from the a range of values
Replicator Neural Networks (RNN) [51]	Provides a measure of outlines of data records
Autoencoder neural networks [52]	Finds anomalies using the dimensional reduction technique
Clustering methods [53]	Once clusters are formed, data not lying in any of the clusters are outliers
Optical Character Recognition (OCR) [54]	Converts scanned documents to text without human intervention, thereby reducing the chances of introducing errors
Random forest [55, 56]	Closes data gaps by accurately predicting missing data over a wide range of datasets
Natural Language Processing (NLP) [57]	Reduces the likelihood of error by automatically identifying which category the answer can be coded to
Clustering methodologies [57]	Cleans biomedical metadata by clustering similar keys together

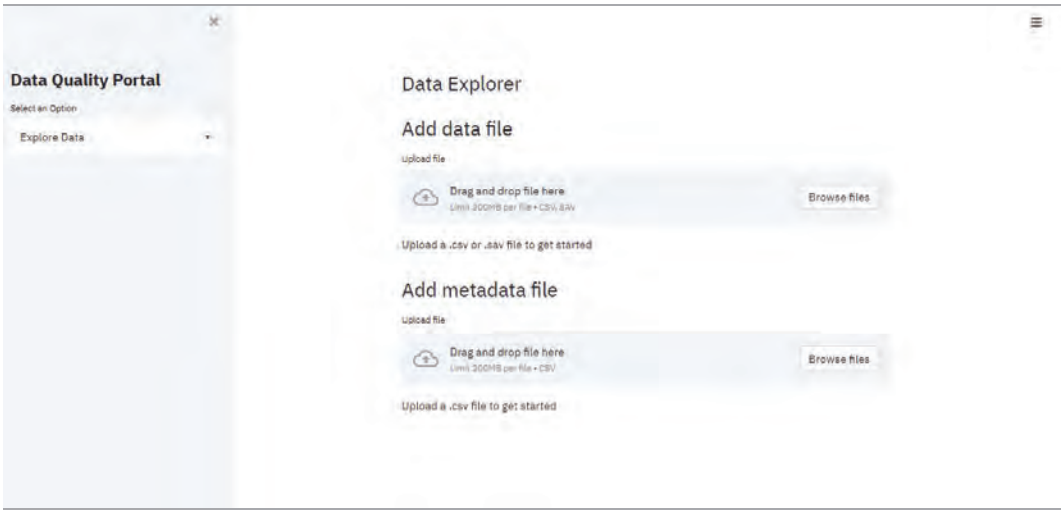
Though these techniques have been suggested to ensure good quality data, it is advisable to first try each of them either on a snippet of the data or on a copy of the original data. Only once reviewed and tested, and the results obtained are convincing and desirable, should one use these above-mentioned machine learning techniques on the whole dataset to have an improved quality data.

Examples of application of machine learning techniques to assess data quality

Two machine learning tools that automate data quality checks and labelling are now available in the public domain.

The Data Quality Label Tool

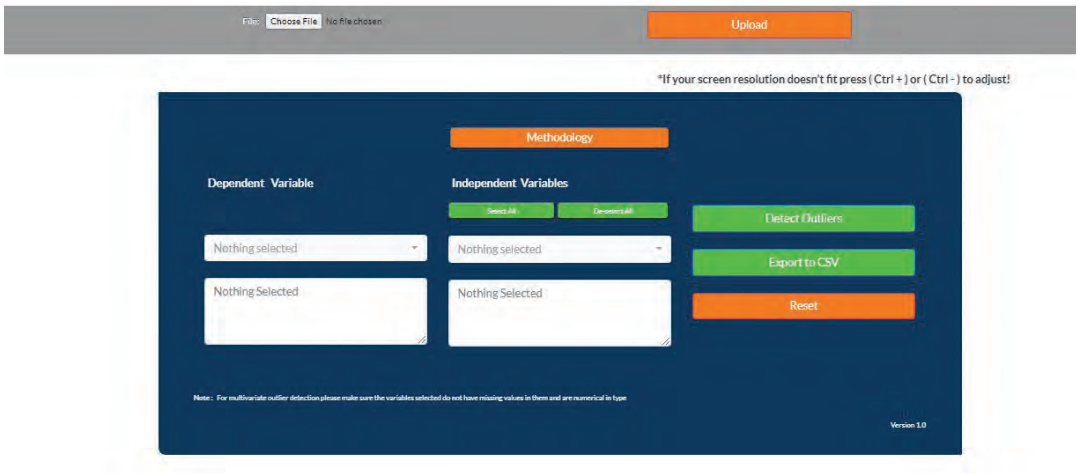
In collaboration with NDQF, and supported by the Population Council, a team of researchers at IIIT-Delhi has developed a tool using ML techniques to measure the data quality. The domain agnostic tool takes a query dataset and its codebook to derive a composite score combining provenance, meta-data coupling, anomalous features, and statistical properties among others. The underlying model was trained on data and meta-data from more than 250 publicly available datasets and validated on multiple rounds of NFHS datasets. The data quality assessment tool is publicly available as a web application at https://dataquality.tavlab.iiitd.edu.in.



Snapshot of The Data Quality Label Tool

Outlier Detection Tool

NDQF data science lab has developed an outlier detection tool to identify the potential outliers in the dataset using machine learning techniques. The tool works for any survey dataset using multiple data science approaches like silhouette score calculations, k-means clustering and isolation forest to flag observations that are potential outliers. Unlike most methods of outlier detection that helps in identifying outliers within one variable (one-dimensional), this tool will help to solve the bigger challenge of finding outliers in multidimensional space. The outlier detection tool is available in public domain (*https://ndqf001.pythonanywhere.com/*).



Snapshot of Outlier Detection Tool

References

1. Lavrakas PJ. Encyclopedia of survey research methods: Sage publications, 2008.
2. Canadian Institute for Health Information. The CIHI Data Quality Framework. Ottawa, Ont., 2009.
3. Health Information and Quality Authority. International Review of Data Quality. 2011.
4. Health Information and Quality Authority. Annual report 2018.
5. United Nations. United Nations National Quality Assurance Frameworks Manual for Official Statistics (UN NQAF Manual) In: Department of Economic and Social Affairs UN, ed., 2019.
6. United Nations Statistics Quality Assurance Framework. UN Statistics Quality Assurance Framework Including a Generic Statistical Quality Assurance Framework for a UN Agency. 2018.
7. Statistics Canada Methodology Branch (SCMB). Statistics Canada Quality Guidelines. Fourth edition. Ottawa, Ontario, Canada K1A 0T6, 2003.
8. European Statistical System (ESS). Quality Assurance Framework of the European Statistical System (Version 2.0). 2019.
9. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey methodology second edition, 2009.
10. Adèr HJ, Mellenbergh GJ. Research Methodology in the Social, Behavioural and Life Sciences: Designs, Models and Methods. Sage, 1999.
11. Gideon L. Handbook of survey methodology for the social sciences. New York: Springer, 2012.
12. Turner AG, Angeles G, Tsui AO, Wilkinson M, Magnani R. Sampling manual for facility surveys for population, maternal health, child health and STD programs in developing countries. 2000.
13. Indian Council of Medical Research. National ethical guidelines for biomedical and health research involving human participants. In: Research ICoM, ed. New Delhi: Indian Council of Medical Research, 2017.
14. Bhutta ZA. Beyond informed consent. Bulletin of the World Health Organization 2004;82:771-7.

15. Spatz ES, Suter LG, George E, Perez M, Curry L, Desai V, Bao H, Geary LL, Herrin J, Lin Z, et al. An instrument for assessing the quality of informed consent documents for elective procedures: development and testing. BMJ open 2020;10(5):e033297.
16. Welch BM, Marshall E, Qanungo S, Aziz A, Laken M, Lenert L, Obeid J. Teleconsent: a novel approach to obtain informed consent for research. Contemporary clinical trials communications 2016 Aug 15;3:74-9.
17. National Drug Abuse Treatment Clinical Trial Network. Internet: https://gcp.nidatraining.org/resources.
18. Guideline IHT. Guideline for good clinical practice. J Postgrad Med 2001;47(3):199-203.
19. UNFPA and the Population Council. Operations Research Methodology Options: Assessing Integration of Sexual and Reproductive Health and HIV Services for Key Affected Populations. August 2013.
20. Kish L. Survey Sampling. New York: John Wiley & Sons, Inc., 1995.
21. Martínez LI. Technical report on Improving the use of GPS, GIS and RS for setting up a master sampling frame. Technical Report Series GO-06-2015: FAO, 2013.
22. Roy TK, Acharya R, Roy A. Statistical Survey Design and Evaluating Impact. New Delhi, London and New York: Cambridge University Press, 2016.
23. Measure DHS. DHS Survey Organization Manual 2012.
24. International Institute for Population Sciences (IIPS) and ICF. National Family Health Survey (NFHS-4), 2015–16: Interviewer's manual, Mumbai: IIPS. Mumbai: IIPS, 2014.
25. Centers for Disease Control and Prevention. Global Adult Tobacco Survey (GATS): Core Questionnaire with Optional Questions, Version 2.0. Atlanta, GA: Global Adult Tobacco Survey Collaborative Group., 2010:2010–56.
26. Macro International Inc. AIDS Indicator Survey: Household Listing Manual MEASURE DHS Calverton. Maryland, USA, 2007.
27. Peterson RA. Constructing effective questionnaires. Thousand Oaks, CA: Sage, 2000.
28. Macro ICF. Training field staff for DHS surveys. Calverton, MD. ICF Macro, 2009:724.
29. World Health Organization. Health in all policies: training manual. World Health Organization. 2015.

30. Measure D. Demographic and health survey sampling and household listing manual. Calverton: ICF International, 2012.
31. Measure D. Demographic and health survey biomarker field manual. Calverton: ICF International, 2012.
32. World Bank. Internet: https://dimewiki.worldbank.org/wiki/Enumerator_Training.
33. Raiten DJ, Namasté S, Brabin B, Combs GJ, L'Abbe MR, Wasantwisut E, Darnton-Hill I. Executive summary--Biomarkers of Nutrition for Development: Building a Consensus. Am J Clin Nutr 2011;94:633S-50S.
34. EURECCA. Internet: www.eurecca.org 2020.
35. National Health and Nutrition Examination Survey. Internet: http://www.cdc.gov/NCHS/NHANES.htm.
36. UNICEF and Population Council. Comprehensive national nutrition survey: anthropometric measurement manual. March 2016.
37. National Health and Nutrition Examination Survey. Anthropometry Procedures Manual,. Centers for Disease Control and Prevention, 2011.
38. World Health Organization. Data quality review: a toolkit for facility data quality assessment. Geneva: World Health Organization, 2017.
39. Ehling M, Körner T. Handbook on data quality assessment methods and tools. European Commission, Eurostat, 2007.
40. Measure evaluation. Internet: https://www.measureevaluation.org.
41. Kish L. Weighting for Unequal P_i. Journal of Official Statistics 1992;8(2):183-200.
42. National Household Survey Capability Programme UN. Household Surveys in Developing and Transition Countries (Studies in Methods. Series F; No 96), 2005.
43. Singh S. A new method of imputation in survey sampling. Statistics 2009;43(5):499-511.
44. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review 2010;78(1):40-64.
45. Kalton G, Kish L. Some efficient random imputation methods. Communications in Statistics - Theory and Methods 1984;13(16): 1919-39.

46 Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence 2019;33(10):913-33.

47. Chew RF, Amer S, Jones K. Residential scene classification for gridded population sampling in developing countries using deep convolutional neural networks on satellite imagery. International Journal of Health Geographics 2018;17(12).

48. Shah N, Mohan D, Bashingwa JJH, Ummer O, Chakraborty A, LeFevre AE. Using Machine Learning to Optimize the Quality of Survey Data: Protocol for a Use Case in India. JMIR Res Protoc 2020;9(8):e17619.

49. Brownlee J. 4 Automatic Outlier Detection Algorithms in Python. 2020.

50. Kumar D. Fraud Analytics using Extended Isolation Forest Algorithm. 2020.

51. Hawkins S, He H, Williams G, Baxter R. Outlier Detection Using Replicator Neural Networks. In: Kambayashi Y., Winiwarter W., Arikawa M. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2002. Lecture Notes in Computer Science,. In: Springer B, Heidelberg, ed., 2002.

52. Fuertes. T. Internet: https://quantdare.com/outliers-detection-with-autoencoder-neural-network/.

53. Ryan M. Internet: https://medium.com/datadriveninvestor/how-to-clustering-and-detect-outlier-at-the-same-time-30576acd75d0.

54. Infosys. Using machine learning in data quality management. In: Infosys, ed., 2020.

55. Kumar S. 7 Ways to Handle Missing Values in Machine Learning. 2020.

56. Gupta A. Internet: https://medium.com/airbnb-engineering/overcoming-missing-values-in-a-random-forest-classifier-7b1fc1fc03ba.

57. Hu W, Zaveri A, Qiu H. Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. BMC Bioinformatics 2017;18:415.

Selected Definitions

- **Analysis of Variance (ANOVA)**
  A statistical technique to analyse variation in a continuous variable measured under conditions defined by categorical variables, often with nominal levels.
- **Artificial Intelligence**
  It is the ability of a computer or a robot controlled by a computer to perform tasks that are done primarily by humans and require human intelligence and judgement.
- **Autocorrelation**
  In time series data, autocorrelation is a measure of strength of the relationship between a time-dependent variable's current value and its past values.
- **Back-check**
  It refers to re-interview by the field supervisor or a data quality monitor of randomly selected sub-sample of respondents who are already interviewed by a field investigator using a small set of factual questions from the survey instrument (or questionnaire) with the intention of checking the accuracy of data collected.
- **Blinded Re-measurement**
  It refers to re-measurement of a subject selected randomly and whose original measurer is not known to the repeat measurer. The purpose of 'blinded re-measurement' is to evaluate the accuracy of original measurement without any bias for or against the original measurer.
- **Calibration**
  It is a process of ensuring and maintaining the accuracy of a measuring instrument in alignment with a standard or accepted range of results.
- **Coefficient of Variation (CV)**
  It is a statistical measure of the dispersion of data points showing the extent of variability of data in a sample in relation to the sample mean. It is calculated as the ratio of the standard deviation to the mean, and is useful for comparing two samples even if the means are different from one another.
- **Cold-deck Imputation**
  It refers to the imputation of missing observations by values from a source unrelated to the data set under consideration.
- **Confidence Intervals**
  It is a range of estimated values in which the true value of a parameter lies with a specified level of certainty.
- **Coverage Error**
  It is the error in an estimate resulting from failure to cover accurately all targeted units of the study population.
- **Dashboard**
  It is a graphical tool used for information management as it organises and displays important metrics into one easy to access place. It provides a quick view of levels and trends in different indicators as well as their interrelationships.
- **Data Security**
  This refers to practice of protecting digital information from unauthorised access, manipulation, or theft.

- **Decision Tree**
  It is a decision support tool that looks like a tree structure and is used for classification and prediction modelling. The uses of decision tree are found in operations research (decision analysis) and machine learning.
- **Equal Probability of Selection Method (EPSEM)**
  EPSEM is a sampling technique that results in the population elements having equal probabilities of being included in the sample.
- **External Validity**
  It refers to how well the outcome of a study can be generalised with respect to different measures, persons, settings, and times.
- **Field Check Table (FCT)**
  These are a set of tools to understand the progress of survey work in field, track any significant departure from expected distributions of important population parameters and identify problematic survey teams or individual investigators as source of bias/ systematic error, if any. FCT is usually generated and discussed at a week or two-week time lag.
- **Haemolysis**
  It refers to the process of the destruction of red blood cells which leads to the release of haemoglobin into the blood plasma.
- **Hot-deck Imputation**
  It refers to imputation of missing observations by values of similar responses from the same data source.
- **Imputation**
  Imputation is a method of estimating and filling in missing values in data.
- **Interquartile Range**
  It is a measure of dispersion in data, computed by taking difference between 75^{th} (3^{rd} quartile) and 25^{th} (1^{st} quartile) percentiles.
- **Isolation Forest**
  It is an unsupervised machine learning algorithm that is used for anomaly detection and works on the principle of isolating anomalies/outliers.
- **K-means Clustering**
  It is a type of unsupervised machine learning algorithm that partitions available observations into several clusters where each observation belongs to the cluster with the nearest mean.
- **Kurtosis**
  It is the measure of tailedness or peakedness of frequency distribution of a variable in data, that is, how tall or sharp the central peak of the distribution is while measured with respect to a normal distribution.
- **Listing**
  It refers to the process of identifying the target population or households for developing a sampling frame.
- **Machine Learning**
  It refers to the study of computer algorithms that improve on its own through experience and the use of data.
- **Mean Value Imputation**
  It is an imputation method where missing values of a variable are replaced with the mean of the non-missing values for the same variable.
- **Measurement Error**
  It is the difference between a measured quantity and its true value.

- **Negative Screening Rate**
  It is defined as ratio of the number of screening questions marked 'No' by the interviewer to the total number of valid screening questions in the instrument.
- **Neural Network**
  Neural network is a computational learning system using a network of functions to understand and translate data into a desired output, usually in a different form.
- **Non-sampling Error**
  It is a type of error, not related to sampling of units, arising during the survey process, such as failure to locate and interview the correct household, asking questions incorrectly, and data entry errors.
- **Optical Character Recognition**
  This is a kind of technology that is used to convert virtually any image containing texts (typed or handwritten or printed) into machine-readable text data.
- **Paradata**
  In a survey, it refers to auxiliary data collected about interviews and survey processes. Some examples of paradata include (not limited to) duration of interview, time taken to ask each question, and negative screening.
- **Parallax Error**
  This is an error caused in reading of a measurement (for example, in anthropometry) due to a viewing angle that is other than an angle perpendicular to the object being measured.
- **Precision**
  It refers to how closely repeated measurements (or observations) of an object (or indicator) come to duplicating measured or observed values.
- **Pre-testing**
  It refers to the stage in survey research when survey questions and questionnaires are tested on members of target population/study population, to evaluate the reliability and validity of the survey instruments prior to start of the survey.
- **Probability Sampling**
  It is a sampling technique in which samples are drawn from the target population using methods based on the theory of probability. Each sample selected in this method has a specified probability of selection.
- **Primary Sampling Units (PSU)**
  It refers to the set of sampling units from where units are selected in the first (primary) stage of a multi-stage sampling design.
- **Random Forest**
  Random forest is an ensemble learning method used for classification, regression and other tasks. It operates by constructing numerous decision trees at training time and displaying the class which is the mode of the classes or the mean/average prediction of the individual trees.
- **Random Imputation**
  It refers to the process where observations of an attribute are drawn randomly from the dataset for imputing the missing values of that attribute.
- **Response Error**
  It represents inaccuracies in responses to questions asked during sample surveys and arises due to a number of reasons including problems with the survey instrument or its implementation and respondent's understanding of the questions.

- **Response Rate**
  It refers to the proportion of sample who responded to a survey out of the total number of target sample.
- **Sampling Design**
  It refers to the methodology used to select sample units for measurement from a specified population and is described by defining sampling universe, sampling frame, stages of sampling and method of sampling at each stage.
- **Sampling Error**
  It is the deviation between a sample estimate and the population parameter under study, caused by sampling design or sample selection.
- **Sampling Frame**
  It refers to the list of the target population units from which samples are drawn for the data collection.
- **Semi-structured Questionnaires**
  It refers to a type of questionnaire which contains both open-ended and closed-ended questions, that is, it allows some questions to have pre-specified answers and some other to have unspecified answers in text form.
- **Simple Random Sampling**
  A simple random sample is a sampling method in which a sample is selected randomly from a specified population in such a way that each member of the population has an exactly equal chance of being selected into the sample.
- **Skewness**
  It is a measure of symmetry (or asymmetry) of the frequency distribution of a variable in data, with respect to its central point.
- **Spot-check**
  Spot-check is a way to ensure data quality in field surveys. It is when senior survey staff physically observes interviewers conducting interviews.
- **Standard Deviation**
  It is a statistical quantity that measures the amount of variation (or dispersion) of a set of values of a particular variable or in other words, it measures how far the values disperse from the mean value of the variable.
- **Structured Questionnaires**
  It refers to a type of questionnaire with questions that allow only a pre-specified set of responses for each question.
- **Support-vector Machines (SVM)**
  These are supervised machine learning models with learning algorithms analysing data for two-group classification problems as well as regressions.
- **Technical Error of Measurement (TEM)**
  It is an accuracy index to present error-margin in anthopometric and captures both inter-rater and intra-rater variability in anthropometric measurements. It is used to evaluate the accuracy of anthropometry measurers during a training session.
- **Total Survey Error**
  It refers to the accumulation of all the errors that arise in the design, collection, processing, and analysis of survey data.
- **Z-Score**
  It is a statistical quantity that gives one an idea of how far from the mean a particular data point is. Given a set of values, it measures the number of standard deviations below or above the sample mean a particular data is.

List of Contributors

ICMR-National Institute of Medical Statistics

Dr. M Vishnu Vardhana Rao	Scientist G & Director
Dr. Damodar Sahu	Scientist F
Dr. Saritha Nair	Scientist E
Dr. Ravendra Kumar Sharma	Scientist E
Dr. Bal Kishan Gulati	Scientist D

Population Council

Dr. Niranjan Saggurti	Director
Dr. Rajib Acharya	Senior Associate
Dr. Bidhubhusan Mahapatra	Project Director
Dr. Sowmya Ramesh	Senior Program Officer
Dr. Nizamuddin Khan	Senior Program Officer
Mr. Akash Porwal	Program Officer
Ms. Trisha Chaudhuri	Data Scientist

NDQF Project Staff

Dr. Yaisna RK	Scientist C
Dr. Vijit Deepani	Scientist B
Ms. Itishree Nayak	Scientist B
Ms. Kanika Sandal	Scientist B
Ms. Priyanka Kakkar	Senior Research Fellow

Notes

Notes

Notes

ICMR-National Institute of Medical Statistics,
Ansari Nagar, New Delhi-110029, India
Phone: +91-11-26588803
Fax: +91-11-26589635
http://icmr-nims.nic.in