Indian J Med Res 156, December 2022, pp 715-720
DOI: 10.4103/ijmr.ijmr_1261_22

Quick Response Code:

Process Paper

National guidelines for data quality in surveys: An overview

M. Vishnu Vardhana Rao[1], Damodar Sahu[1], Saritha Nair[1], Ravendra Kumar Sharma[1], Bal Kishan Gulati[1], Rajib Acharya[2], Bidhubhusan Mahapatra[2], Sowmya Ramesh[2], Nizamuddin Khan[2], Trisha Chaudhuri[2], Kanika Sandal[1], Vijit Deepani[1], Sangeeta Dey[1] & Niranjan Saggurti[2]

[1]ICMR-National Institute of Medical Statistics & [2]Population Council India, India Habitat Centre, New Delhi, India

Received June 5, 2022

Good quality health, nutrition and demographic survey data are vital for evidence-based decision-making. Existing literature indicates system specific, data collection and reporting gaps that affect quality of health, nutrition and demographic survey data, thereby affecting its usability and relevance. To mitigate these, the National Data Quality Forum (NDQF), under the Indian Council of Medical Research (ICMR) - National Institute of Medical Statistics (NIMS) developed the National Guidelines for Data Quality in Surveys delineating assurance mechanisms to generate standard quality data in surveys. The present article highlights the principles from the guidelines for informing survey researchers/organizations in generating good quality survey data. It describes the process of development of the national guidelines, principles for each of the survey phases listed in the document and applicability of them to data user for ensuring data quality. The guidelines may be useful to a broad-spectrum of audience such as data producers from government and non-government organizations, policy makers, research institutions, as well as individual researchers, thereby playing a vital role in improving quality of health, nutrition and demographic data ecosystem.

Key words Data quality - health survey - quality assurance

Need for data quality assurance

India is a data-rich country with data being generated through multiple sources used to inform the policies and programmes. Despite the availability of rich data, utilization is often limited owing to quality issues[1,2]. Existing literature suggests three main data quality (DQ) challenges specifically in health, nutrition and demographic surveys. These include: data collection issues: length of the questionnaire and related bias, data collectors' behavioural bias; data entry challenges and lack of accountability[3]; reported data issues: divergent demographic numbers/estimates for the same indicator from different sources; inconsistency within data sources and incomplete or missing data[4-6]; system-specific issues: lack of standardized questions/ indicators, lack of DQ assurance mechanisms and audits, lack of trainings on the value of data and limited use of technology[7].

The quality of training and interviewing skills of field officers also has an impact on the quality of data[8].

Poor quality data not only impacts the decision-making processes and programme planning but also affects the organizational culture, timely intervention and operational costs[9-14]. Although each survey has its own DQ checks and balances, currently, there is limited DQ assurance guidance for surveys that can be adopted by individuals/agencies/institutes/organizations for ensuring high-quality, reliable data.

The National Data Quality Forum (NDQF), a joint venture of the ICMR-NIMS and Population Council, India, envisions to improve the quality of the health and demographic data ecosystem in India. The Forum specifically aims to educate and deepen interest among producers and users for the highest standard and congregate them to a specific platform to facilitate information sharing as well as equip them with appropriate solutions, guidance, strategies and tools to contribute to the improvement in DQ. Against this backdrop, NDQF developed the National Guidelines for Data Quality in Surveys. This manuscript describes the process of development of the national guidelines and highlights the principles from the guidelines to inform survey researchers/organizations regarding generation of good quality data. The goal of the national guidelines is to outline a comprehensive list of principles to be followed while designing and implementing health, nutrition and demographic surveys to mitigate the errors and biases that may creep at various stages, including designing, planning, data collection and analysis. In addition, the guidelines aim to sensitize users and producers of survey data about best practices in DQ, help in developing strategies and institutionalizing DQ assurance mechanisms.

Development of national guidelines

The development of the aforementioned guidelines began with an extensive review of national and international literature[8,15-19] on DQ frameworks and dimensions adopted by institutes/agencies to ensure quality data collection. The literature review emphasized the need to establish effective and efficient quality assurance protocols towards improvement in the overall survey DQ.

Following this, a series of consultations were held with experts engaged in planning and implementation of surveys in India. Literature review and deliberations with experts resulted in the drafting of a comprehensive guideline by NDQF core team. The guidelines, applicable for in-person quantitative surveys, enlist core components and key measures to be practiced in each of the three phases – preparatory (pre-data collection); data collection and post-data collection phase.

Again, a series of consultative meetings were held with experts and the draft guideline was revised incorporating the feedback. The draft was critically reviewed by the technical advisory group of the NDQF consisting of eminent scientists and programme personnel and revised. In addition, it was reviewed by the subject experts engaged in population, health and nutrition surveys and finalized. The final version of the guidelines broadly consists of DQ assurance principles and dimensions; sources of errors and biases; ethics and DQ; DQ assurance management plan; DQ assurance systems – pre, during and post-data collection for survey, anthropometry and biological components, checklists to ensure DQ at each stage of the survey, technological tips and machine learning (ML) techniques at each stage of survey. The National Guidelines for Data Quality in Surveys can be accessed at *https://main.icmr.nic.in/sites/default/files/upload_documents/National_Guidelines_for_DATA_QUALITY_in_Surveys.pdf or https://ndqf.in/guidelines/*.

Overview

The National Guidelines provide insights on essential steps to be adhered from the designing stage of the survey to data analysis stage for ensuring good quality data. Based on the phases of a survey, the National Guideline is categorized into three sections: preparatory (before starting data collection); during and post-data collection. Each section highlights the points to be considered for survey designing, implementation and analysis. It provides a comprehensive list of tools and frameworks that can be used as it is or adapted to a different context. The document only provides guidance on quality assurance and not the technicalities of survey components. In addition, the guideline provides essential checklists and technological tips that can be used to enhance DQ. The details of each section of the guidelines are presented in Figure.

Data quality assurance (DQA) monitoring: Adhering to quality assurance protocols during the planning phase of a survey is important. This includes examining the appropriateness of the study design, understanding the provenance of error and bias at each stage of survey, performing quality assurance activities through planned field visits, analytics on crucial data items and reporting measures on the quality of collected data.

A comprehensive quality assessment framework, essential to ensure good quality data consists of three

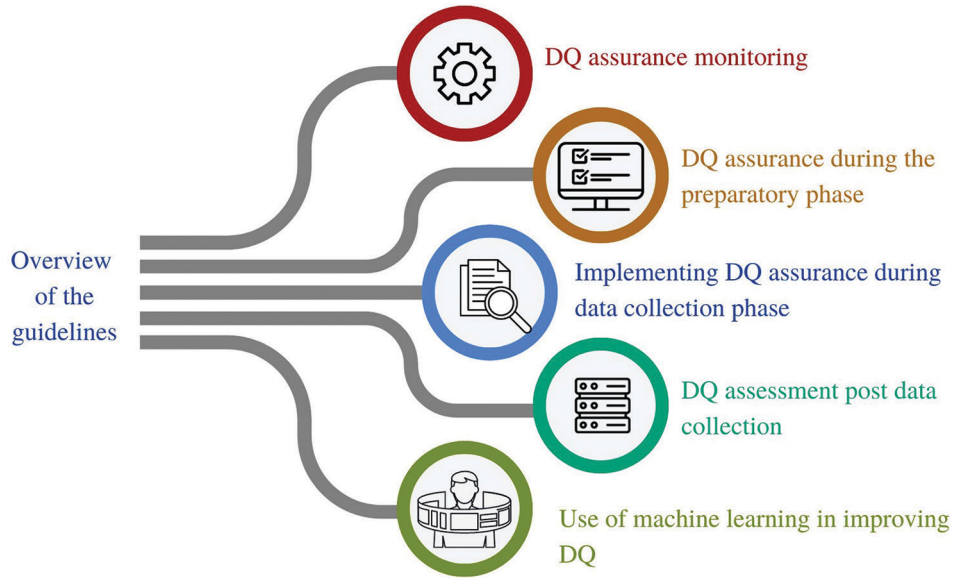

Figure. Overview of the national guidelines. DQ, data quality

principle domains – survey institution (organizational and institutional dimensions); survey setting (quality dimensions to be followed at the time of data collection – methodological soundness; data confidentiality; response burden and cost-efficiency) and survey data product (quality dimensions to be adhered to once data has been collected – relevance; accuracy; reliability; accessibility/clarity; coherence; comparability; completeness and validity). These dimensions are inter-related and overlapping.

Setting up of an institutional structure to implement DQ management plan, establishing procedures to monitor the implementation of DQ management plan and ensuring compliance to ethical principles are the key domains of DQ assurance mechanisms.

DQA during preparatory phase: The key principles to follow in the preparatory phase include adopting pre-tested sampling frame, assessing the quality of the study design and sampling methodology, quality considerations while developing survey tools, manuals, and data entry applications, reviewing various dimensions of the survey tools to avoid bias and redundancy, quality consideration when recruiting and training survey investigators and ascertaining quality assurance parameters to be implemented for anthropometry and biomarker data collection.

Quality consideration during data collection phase: The quality assurance protocols to be followed during survey data collection phase include laying emphasis on monitoring (anthropometric and biomarker) data collection through regular field visits, developing interactive DQ dashboards for monitoring of data collection in the field setting, performing analytics on critical indicators to evaluate the pattern of data being collected, conducting regular data review meetings, providing feedback to the team and documenting all quality monitoring processes and actions. Survey team members should be provided with an outline of their roles and responsibilities, standard protocols, tools and manuals. For ensuring DQ, regular monitoring with the use of field monitoring checklist, re-interview tools (back-checks and spot-checks), field check tables (update about field progress), para-data (auxiliary data describing data collection processes help in reviewing team movement plan, interviewer-specific indicators and quality of interview) are suggested.

Data quality (DQ) assessments in post data collection phase: During post-data collection phase, various analytical techniques can be used for processing and profiling survey data (undertaking preparatory and descriptive analysis to assess missing values, outliers, and data inconsistencies, preparing complete documentation covering important details of the survey instruments and tools, protocol information and analysis approach) and computing sampling and non-sampling errors and reporting DQ measures on key indicators.

Role of machine learning (ML) in improving quality of survey data: ML techniques help to improve

data monitoring on real-time basis by delineating data collection trends. It encompasses a range of computational techniques, which help in identifying patterns, classifying and detecting outliers in the data, taking into account supervised or unsupervised techniques[20]. In this regard, various ML techniques – such as Convolutional Neural Networks, Decision Trees and Clustering methodologies – can be used based on motives, needs and phases of survey research.

Dissemination of national guidelines

The national guidelines were disseminated through a series of two-day training of trainers (ToT) in workshop mode conducted for senior/mid-level officials to sensitize the participants on a comprehensive list of guiding principles to be followed for DQ assurance and to encourage them to train individuals in their respective organizations for ensuring DQ during surveys. Dissemination was conducted through brief presentations, videos, interactive quiz and question-answer sessions. A total of seven ToT workshops have been conducted in the five regions of India to disseminate the guidelines among 160 trainers from 27 states affiliated to ICMR institutes, academic and research organizations, medical colleges, non-governmental organizations, population resource centres and Ministry of Health and Family Welfare, Government of India.

Conclusions

The national guidelines provide a comprehensive list of principles for ensuring DQ in health, nutrition and demographic surveys. This would be useful for data producers and users engaged in private and government sectors, including research and academic institutions, national and regional level policy-makers and implementers, ministries, non-profit organizations and survey agencies. Adopting these guidelines may strengthen the health, nutrition and demographic data ecosystem, thereby leading to informed decision-making and evidence-based policies.

***Acknowledgment*:** Authors acknowledge survey and subject experts and NDQF Technical Advisory Group who reviewed the guidelines and provided their comments to finalize the same.

***Financial support & sponsorship*:** None.

***Conflicts of Interest*:** None.

References

1. Pandey A, Roy N, Bhawsar R, Mishra RM. Health information system in India: Issues of data availability and quality. *Demogr India* 2010; *39* : 111-28.
2. Mishra A, Vasisht I, Kauser A, Thiagarajan S, Mairembam DS. Determinants of health management information systems performance: Lessons from a district level assessment. *BMC Proc* 2012; *6* (Suppl 5) : O17.
3. James KS, Rajan IS. Third national family health survey in India: Issues, problems and prospects. *Econ Polit Wkly* 2008; *43* : 33-8.
4. Phillips BS, Singhal S, Mishra S, Kajal F, Cotter SY, Sudhinaraset M. Evaluating concordance between government administrative data and externally collected data among high-volume government health facilities in Uttar Pradesh, India. *Glob Health Action* 2019; *12* : 1619155.
5. Gupta M, Rao C, Lakshmi PVM, Prinja S, Kumar R. Estimating mortality using data from civil registration: A cross-sectional study in India. *Bull World Health Organ* 2016; *94* : 10-21.
6. Mahapatra P, Chalapati Rao PV. Cause of death reporting systems in India: A performance analysis. *Natl Med J India* 2001; *14* : 154-62.
7. Sharma A, Rana SK, Prinja S, Kumar R. Quality of health management information system for maternal & child health care in Haryana state, India. *PLoS One* 2016; *11* : e0148449.
8. Üstun TB, Chatterji S, Mechbal A, Murray C. Quality assurance in surveys: Standards, guidelines and procedures. In: *Household sample surveys in developing and transition countries*. Geneva, Switzerland: World Health Organization; 2005. p. 199-230.
9. Wetherill O, Lee CW, Dietz V. Root causes of poor immunisation data quality and proven interventions: A systematic literature review. *Ann Infect Dis Epidemiol* 2017; *2* : 1012.
10. Haug A, Zachariassen F, Van Liempd D. The costs of poor data quality. *J Ind Eng Manag* 2011; *4* : 168-93.
11. Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. *IEEE Trans Knowl Data Eng* 1995; *7* : 623-40.
12. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. *J Manag Inf Syst* 1996; *12* : 5-33.
13. Lee Y, Pipino L, Funk J, Wang RY. *Journey to data quality*. Cambridge, Mass: The MIT Press; 2006.
14. Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM* 2002; *45* : 211-8.
15. Statistics Canada Methodology Branch. *Statistics Canada quality guidelines*, 4th ed. Ottawa, Ontario, Canada: SCMB; 2003.
16. Global Adult Tobacco Survey Collaborative Group. *Global Adult Tobacco Survey (GATS): Quality assurance: guidelines and documentation*. Atlanta, GA: CDC; 2020.
17. International Institute of Population Sciences (IIPS). *Data quality assurance and quality control mechanisms 2019-20. National Family Health Survey 2019-20 NFHS-5*. Mumbai: IIPS; 2019-20.

18. Biemer PP, Lyberg LE. *Introduction to survey quality*. New York: John Wiley & Sons; 2003.
19. Groves RM, Heeringa SG. Responsive design for household surveys: Tools for actively controlling survey errors and costs. *J R Stat Soc A* 2006; *169* : 439-57.
20. 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* : e17619.

For correspondence: Dr M. Vishnu Vardhana Rao, ICMR-National Institute of Medical Statistics, New Delhi 110 029, India
e-mail: dr_vishnurao@yahoo.com