Improving Survey Data Quality using AI: Bihar Case Study

To ensure high-quality, reliable data from a large-scale household survey, we implemented a multi-tiered Quality Assurance (QA) and Quality Control (QC) framework during a digital access and use survey in 10 districts of Bihar, India. The goal was to minimize interviewer-related errors, detect fabricated data (curb-stoning), and deliver a clean, analysis-ready dataset. This briefer describes the importance of survey data quality in low- and middle-income countries (LMICs), outlines our AI-driven QA/QC system, and presents its methods, results, and implications.

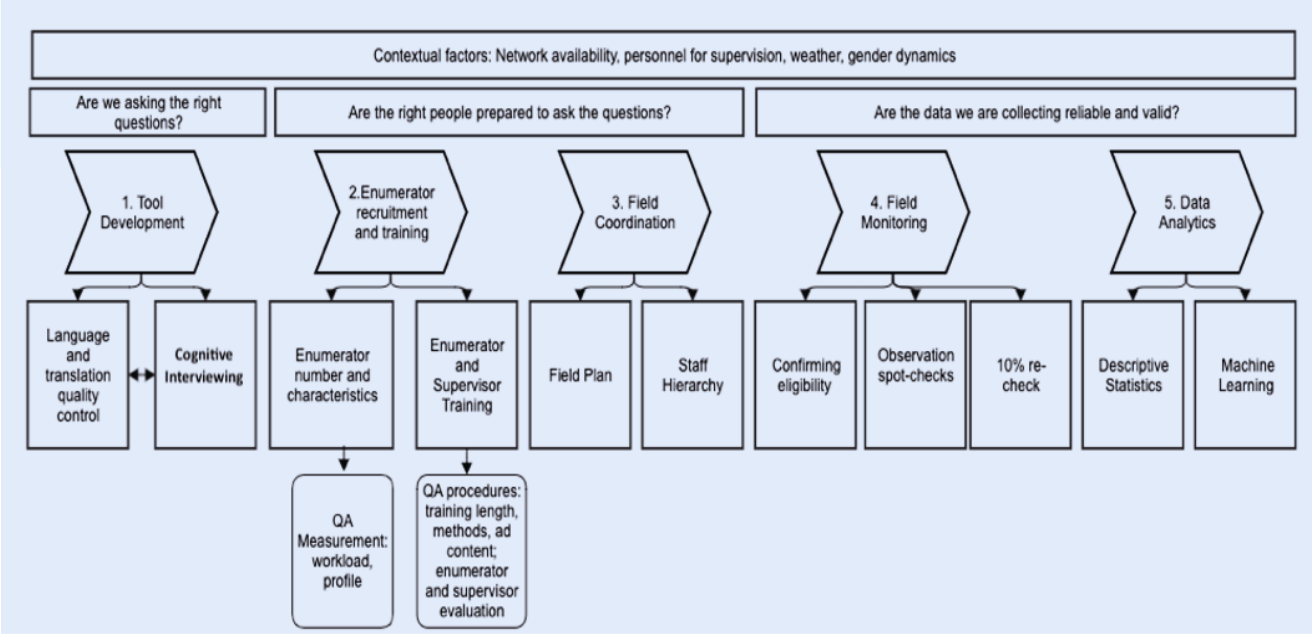
Why is survey data quality important ?

High-quality data are essential for optimizing decision-making and resource allocations in public health and development programming particularly in low- and middle-income countries (LMICs). Special surveys are a critical source of population based health information. However, a range of factors during survey design (question phrasing and translation) and implementation (enumerator error) may impede data quality.

Methods to improve the quality of survey data

Figure 1 depicts our comprehensive framework for improving data quality in large household surveys [1]. To support data collection efforts in Bihar from December 2024 to March 2025, we utilised a machine learning based approach (Step 5. Data Analytics) to produce a similar pipeline for assessing data quality and feeding back results in near real-time to field teams.

Figure 1. Methods to improve the quality of survey data



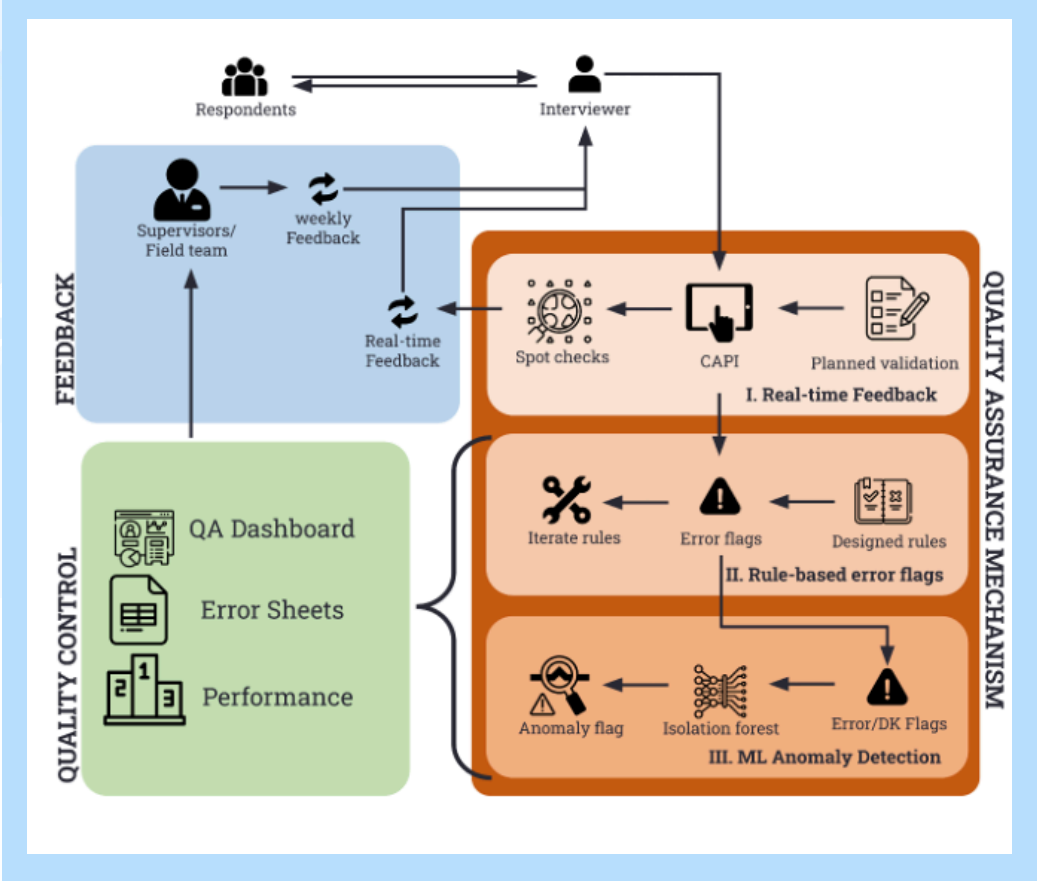
www.evidence-digital.org

Johns Hopkins Bloomberg School of Public Health · School of Public Health Departement Openbare Gesondheid Isikolo Sempilo Yoluntu · University of Cape Town · GxD hub Digital Equity. Connectivity. Insights. · Oxford Policy Management · York University

4-Step quality improvement process

To bolster population-based survey (n = 13,568) data quality in Bihar, a 4-step process was undertaken (Figure 2). This system integrated upfront tool safeguards, continuous monitoring, automated error detection (including machine learning), and feedback loops to the field team. It was complemented by in-person spot checks (re-interviewing ~10% of respondents), and random audits of interviewer behaviour and responses. These on-site verifications provided immediate quality assurance and allowed for the early detection of discrepancies or suspicious patterns during interviews. In the section to follow, we outline the added 5-Step quality improvement process.

Figure 2. Deep dive into field monitoring and data analytics part of quality control system in Bihar



Step 1 During tool development, build safeguards into the CAPI system

Quantitative survey questions were developed following a rigorous phase of qualitative research, including cognitive interviews, which sought to improve survey content and translation. As part of CAPI programming, all questions were assessed for a logical skip framework which was further enhanced during pilot testing. Time stamps were placed at multiple places within the survey tool (e.g. start / stop of modules, start and completion of the interview) with the broader aim of allowing for the tracking for enumerator engagement with the digitized survey tool. Logical safeguards were additionally included throughout the survey tool on questions such as age, date of birth, self-help group membership, and age at first purchase of mobile phone. Safeguards sought to restrict the range of responses provided to those which would be reasonably plausible (e.g. age was restricted to 0 and 100 years of age).

Step 2 Assess data quality and run the data through rule-based error flags and machine learning isolation forest

During data collection, measurable and unmeasurable errors were flagged through rule-based error flags and machine learning. Measurable errors are those that can be quantified, detected, and expressed numerically. These included real-time feedback from the CAPI system itself and from rule-based error flags which sought to identify logical inconsistencies, implausible responses, skip logic failures, as well as newly identified errors derived iteratively throughout the process of data collection. To detect unmeasurable errors, a machine learning isolated forest algorithm was implemented. This sought to explore patterns in the use of the response option of "Don't know" and skipped questions to identify cases where enumerator inputs differed substantially from the norm.

Measurable errors:

1) Real-time feedback:

Automated checks were built into the CAPI software which could validate responses and question administration for skip logic enforcement, range validations, data type enforcement to prevent many errors at entry. This design-driven oversight was implemented during the survey design process and complements traditional field supervision. Some validations were intentionally omitted to assess true interviewer engagement.

2) Rule-Based Error Flag:

Rules base error flags were created to identify and capture various response errors based on:

- Logical inconsistencies (e.g. contradictory SHG participation)
- Implausible responses (e.g. age at mobile ownership <10)
- Skip logic failures or missing critical fields
- Iterative rules updates: New flags introduced during data collection based on observed errors

Unmeasurable errors: Certain errors could not be quantifiably measured such as over reporting don't-knows or fabricated data. For this we used an AI algorithm.

3)Machine Learning: Isolation Forest [2,3]

- Objective: Detect subtle data quality issues missed by rule-based checks
- Isolation Forest algorithm identifies data which shows distinctly different response patterns based on input parameters. Identifies records which are anomalous compared to other records
- Top 5% of anomalous records flagged
- "Don't know" patterns in simple questions (e.g. phone lock) revealed inattentive or falsified entries

Step 3 Generate a report and report findings to field team

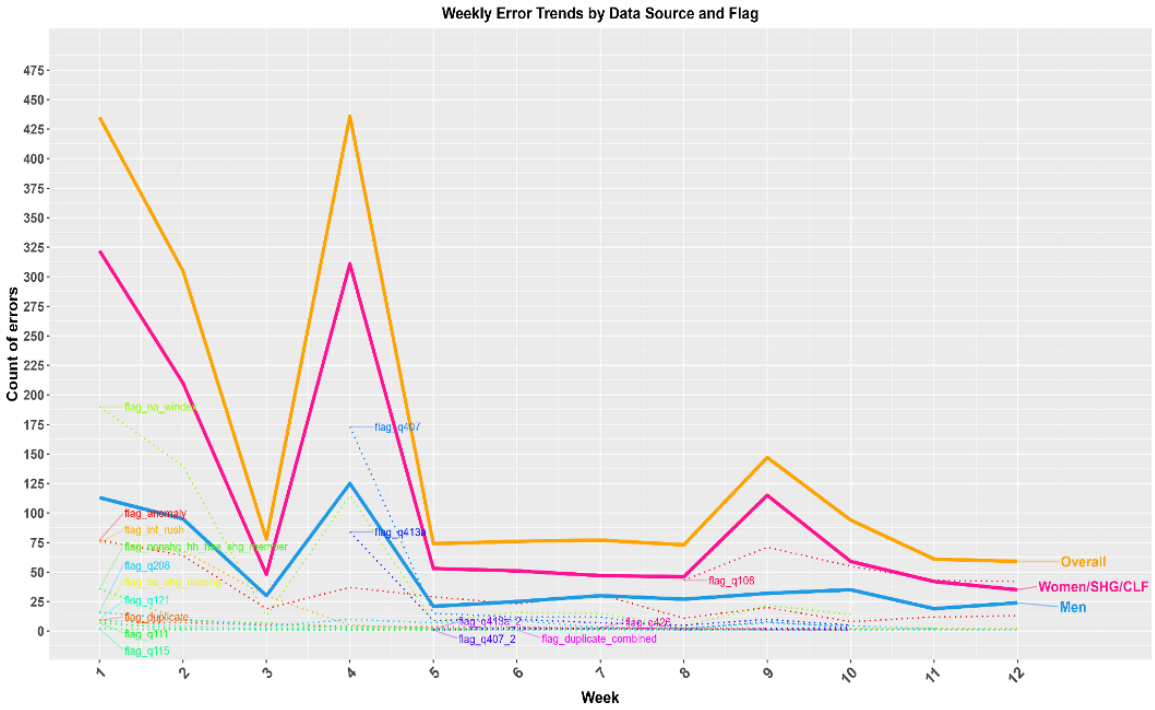
Weekly calls were held amongst the lead data scientist and larger team of senior investigators to review the data quality dashboard and errors flagged throughout the process of data collection. In addition, weekly error reports were shared with the survey team and field team in an automated dashboard format. Excel sheets were generated which listed case-by-case errors, tracked individual enumerator and supervisor performance and provided a detailed overview of each case and critical variables that triggered each error. The field team reviewed these error sheets on a weekly basis with the enumerators.

In the future, more individualised messages might be considered. In prior evaluations in India, our team has done this with success via SMS feedback loops and email alerts to relay data quality issues to field staff in real time [4]. This work described an SMS-based QA system that improved data accuracy. In the Bihar survey, we opted not to use automated SMS due to the survey's complexity and scale which would have made SMS alerts costly and potentially overwhelming for field staff.

Step 4 Field team takes corrective action

The field teams received feedback and supervisors sought to address concerns with data quality in near real-time. Respondents were recontacted where feasible and needed to clarify responses. The impact of this intervention is seen in Figure 3. Where weekly error counts reduced by 85% from the start of the survey (Over 420) within 8 weeks (under 60).

Figure 2. Weekly error tracking and quality assessment in Bihar survey



References

1. *Shah N, Mohan D, Bashingwa J, Ummer O, Chakraborty A, LeFevre A, Using Machine Learning to Optimize the Quality of Survey Data: Protocol for a Use Case in India JMIR Res Protoc 2020;9(8):e17619. DOI: 10.2196/17619*
2. *Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining [Internet]. 2008 [cited 2025 Mar 27]. p. 413–22. Available from: https://ieeexplore.ieee.org/document/4781136*
3. *Lim Y. Unsupervised Outlier Detection with Isolation Forest [Internet]. Medium. 2022 [cited 2025 Mar 27]. Available from: https://medium.com/@limyenwee_19946/unsupervised-outlier-detection-with-isolation-forest-eab398c593b2*
4. *Shah N, Ummer O, Scott K, Bashingwa JJH, Penugonda N, Chakraborty A, Sahore A, Mohan D, LeFevre AE; Kilkari Impact Evaluation Team. SMS feedback system as a quality assurance mechanism: experience from a household survey in rural India. BMJ Glob Health. 2021 Jul;6(Suppl 5):e005287. doi: 10.1136/bmjgh-2021-005287. PMID: 34312150; PMCID: PMC8728370.*

For more information or permission to adapt this resource, please contact:

Dr. Amnesty Lefevre
Director, Evidence for Digital Transformation (EDiT) Consortium
aelefevre@gmail.com

This work was supported by the Gates Foundation.

EDiT EVIDENCE FOR DIGITAL TRANSFORMATION Consortium

www.evidence-digital.org

Consortium partners

Johns Hopkins Bloomberg School of Public Health

School of Public Health, Departement Openbare Gesondheid, Isikolo Sempilo Yoluntu, University of Cape Town

GxDhub Digital Equity. Connectivity. Insights.

Oxford Policy Management

York University