24 September 2025

13th Meeting of the Expert Group on ICT Household Indicators (EGH), Geneva, Switzerland

Real-time quality assurance in ICT household surveys

Dr. Mayank Date, *BDS, MPH*

***Data Scientist**, Johns Hopkins Bloomberg School of Public Health, USA*

Evidence for Digital Transformation (EDiT) Consortium

School of Public Health
Departement Openbare Gesondheid
Isikolo Sempilo Yoluntu
UNIVERSITY OF CAPE TOWN

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Project funded by
Gates Foundation

Data quality assessment processes

Quality Assurance (QA)

Processes that ensure adherence to protocols and early error detection.

Quality Control (QC)

Techniques used to monitor and verify data

Quality Improvement (QI)

Proactive, system-level approach for data-driven efforts that enhance QA and QC systems

Ref: Pavlovic A. https://www.qualio.com/blog/quality-management-vs-quality-control-vs-quality-assurance

Why do we need high quality data on digital access and use?

- **Paucity of high quality data** on digital access and use, particularly in low resource settings where needed is greatest
- **Data provide important insights into the role technology plays in catalysing health and development**
- **Strategies to improve data quality need to start at survey design and continue through during survey implementation**

5 factors that influence survey data quality

Contextual factors: Network availability, personnel for supervision, weather, gender dynamics

Are we asking the right questions?

1. Tool Development

Cognitive Interviews and Pilot testing

Are the right people prepared to ask the questions?

2. Enumerator recruitment and training

3. Field Coordination

Are the data we are collecting reliable and valid?

4. Field Monitoring

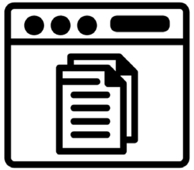
5. Data Analytics

Implementation of AI/ML

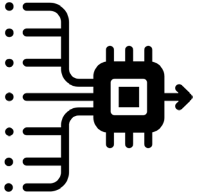
Data Quality Assessment → Iteratively feedback results to inform training, supervision, and ongoing implementation

Digitization of data collection presents opportunities to bolster data quality

- Allow for recording and analysis of **meta and paradata** to identify curbstoning and fabrication



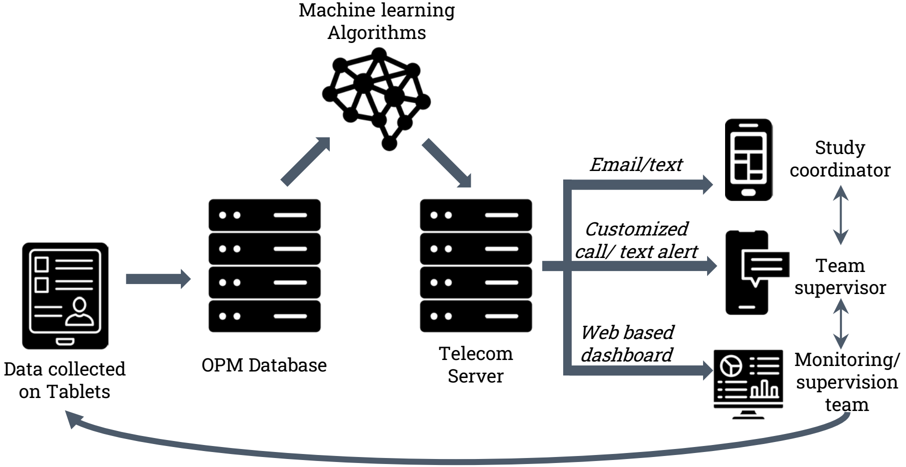
- **Machine learning and AI can be used for error detection** to improve accuracy and consistency in data
    - Examples of implementation in the pharmaceutical and EHR space
    - Limited use in population surveys which represents a missed opportunity



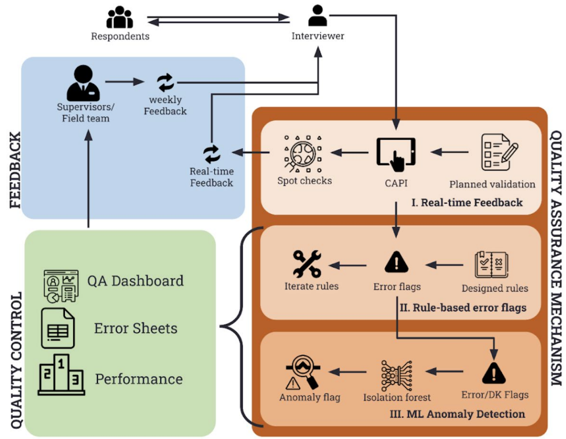
Cohen and Warner - https://www.cambridge.org/core/journals/political-analysis/article/how-to-get-better-survey-data-more-efficiently/6BA27F34CDE007EA8D1AD6692F2E8E60
https://www.sciencedirect.com/science/article/pii/S266731852500011X
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1621514/full
https://www.xiahepublishing.com/m/2771-165X/JCTP-2023-00061

Spotlight on efforts to bolster survey data quality in India

Impact evaluation 2018-2021



Digital Access and Use survey in Bihar, India 2024-2025



Digital access and use survey in Bihar

Population, 2025

Bihar is second most populous state

with 9.2% of national population [Statistics Times]

Gross State Domestic Product, 2024-25

Bihar ranked 14/36

at ₹ 9.76 lakh crore ($ 110 billion) [Forbes]

Human Development Index, 2022

Bihar ranked lowest

of all Indian States [Global Data Lab]

Gender Inequality Index, 2017-18

Bihar ranked 25/36

of all Indian States [Rural India - Working paper]

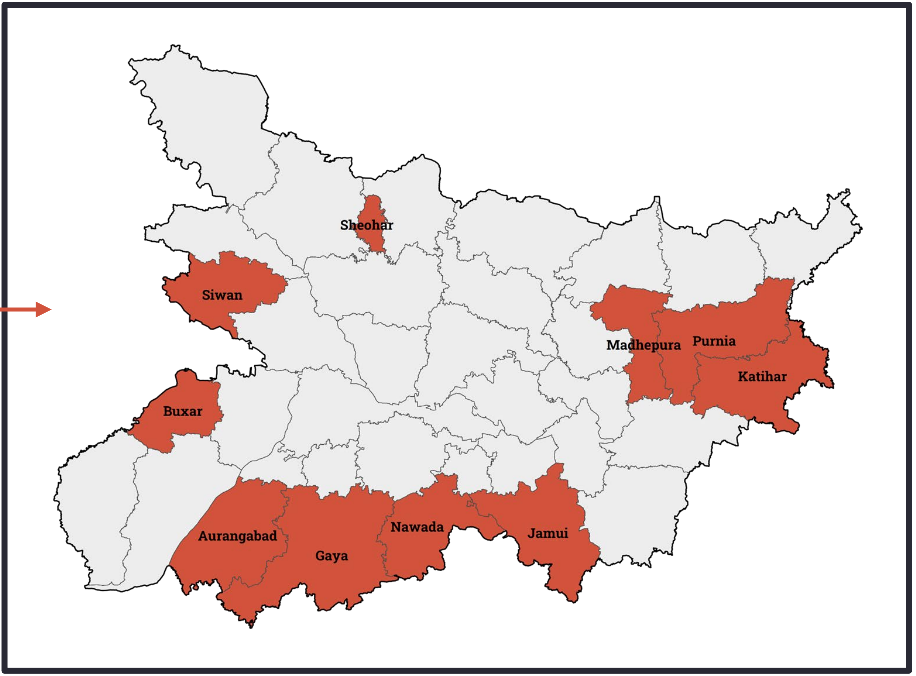
Digital access and use survey in Bihar

Survey aim: Measure population level access to and use of mobile phones among men and women 18-60 years of age across 10 districts of Bihar

Data collection: December 2024-March 2025

Study sites: 10 districts in Bihar

- **>68,000 households** listed across 300 villages
- **13,568 respondents** 18-60 years of age including 8567 women and 5,001 men



Survey Team

- **57 interviewers** from across India; average of 7.5 years of field experience conducting surveys
- **13 supervisors**
- **5 coordinators**

Challenges in data collection and QA/QC

- **Poor network and connectivity –** Teams often faced technical challenges like network issues and power outages

- **Low literate population** – concepts assessed not always familiar to respondents

- **Length of survey** – On average, each interview took about **55 minutes** to complete **Target was 3-5 interviews per day**.
  - Poor respondent engagement, especially with no tangible benefits to them
  - Barriers with understanding local dialects and languages
  - Frequent interruptions from family/children
  - Lack of trust

- **Field conditions–** hot climate, travel distances on foot and by car to reach respondents

- Strict **data storage and sharing policies** prevented implementation of more advanced third-party tools for QA/QC

Steps for setting up QA/ QC Systems

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

- Assess questions for logical skip patterns
- Place time stamps throughout the tool (e.g. start/ stop of sections)
- Assess individual items for logical responses (e.g. age range within 1-100 years)

Quality Assurance

Step 2. Develop rule-based error flags and machine learning algorithm protocols

- Select thresholds and rules for error flags
- Edit rules iteratively during data collection

Steps for setting up QA/ QC Systems

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

- Assess questions for logical skip patterns
- Place time stamps throughout the tool (e.g. start/ stop of sections)
- Assess individual items for logical responses (e.g. age range within 1-100 years)

Step 2. Develop rule-based error flags and machine learning algorithm protocols

- Select thresholds and rules for error flags
- Edit rules iteratively during data collection

Quality Assurance

Step 3. Run data check regularly, track errors and performance

Step 4. Generate error reports

Quality Control

Steps for setting up QA/ QC Systems

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

- Assess questions for logical skip patterns
- Place time stamps throughout the tool (e.g. start/ stop of sections)
- Assess individual items for logical responses (e.g. age range within 1-100 years)

Step 2. Develop rule-based error flags and machine learning algorithm protocols

- Select thresholds and rules for error flags
- Edit rules iteratively during data collection

Quality Assurance

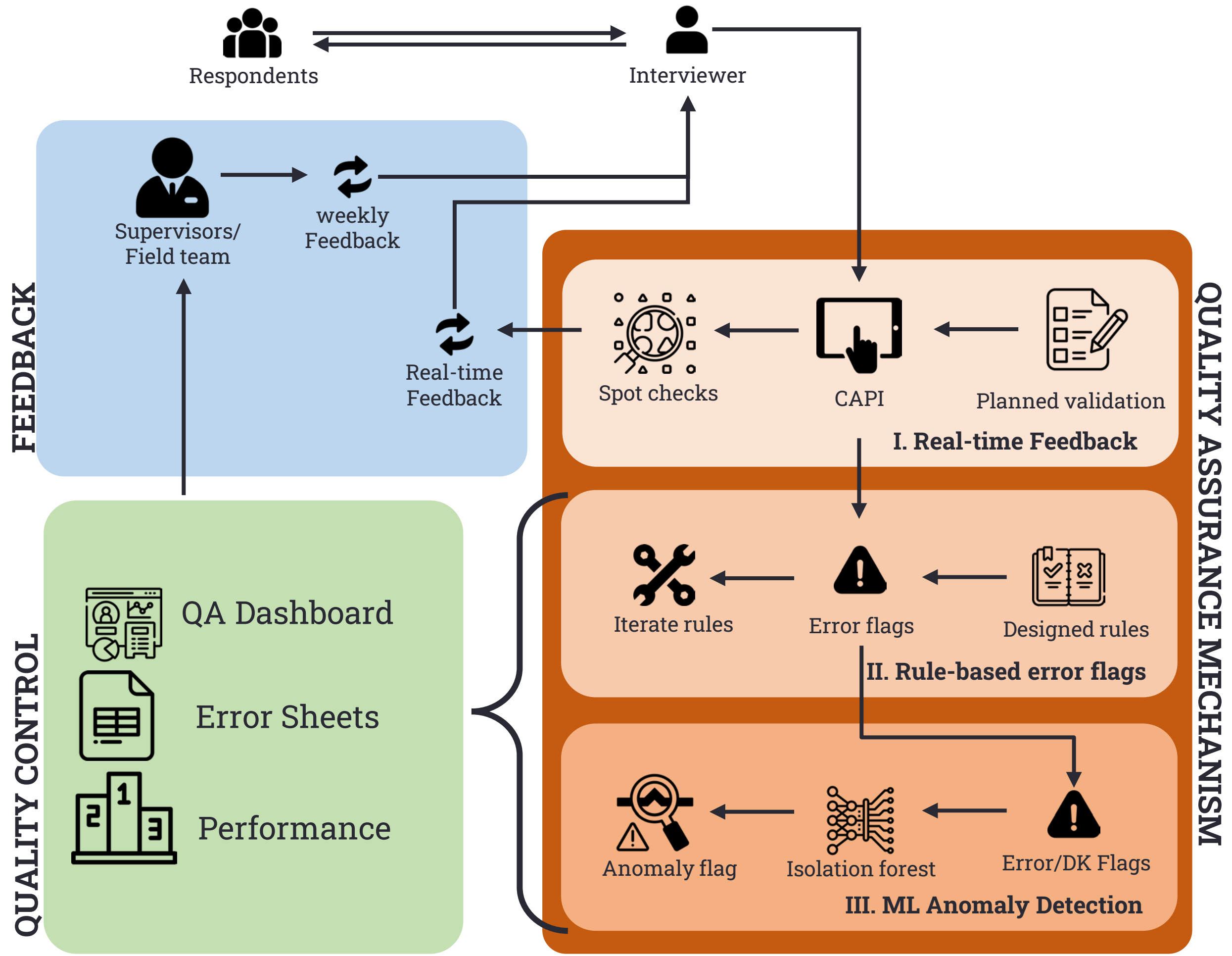
Step 3. Run data check regularly, track errors and performance
Step 4. Generate error reports

Quality Control

Step 5. Share error sheets with field team for corrective action

Feedback

QA/QC framework in practice



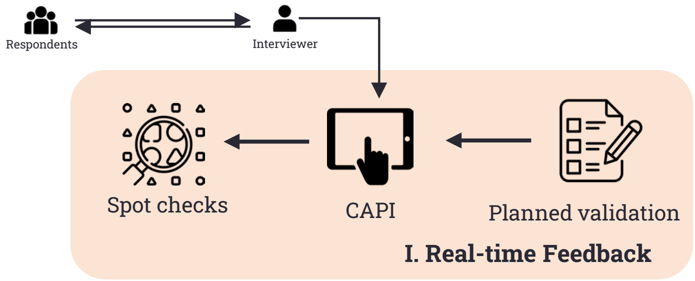
I. Real-Time Feedback

Q425 When was the mobile phone within your reach yesterday? In the morning, in the afternoon, in the evening, or in the night?

```
I [Interviewer: Select all that apply. If the phone was
with the respondent for even a part of the time period,
include that time period in the response]
V1 !(self.Contains(6) && self.ContainsAny(1,2,3,4,5))
M1 Not at all can not be select with any other option
V2 !(self.Contains(1) && self.ContainsAny(2,3,4,5))
M2 Whole Day cannot be select with any other option
```

MULTI-SELECT

01 ☐ Whole day
02 ☐ in the morning (6am - 12pm)
03 ☐ in the afternoon (12pm - 6 pm)
04 ☐ in the evening (6pm - 10pm)
05 ☐ in the night (10pm - 6 am)
06 ☐ Not at all



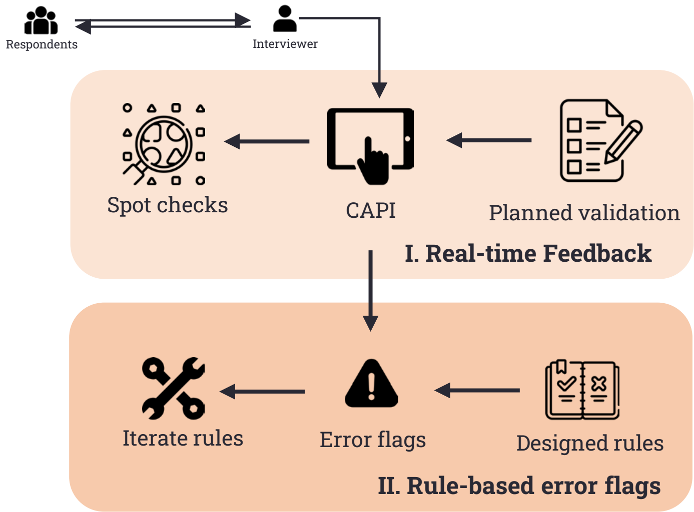
- **Validations programmed into the CAPI**
  software ensuring only valid responses would be accepted
  - Certain age ranges
  - Data format (Numeric/character)
  - Logical selection in multi select questions

- **Spot Checks**
  - About 10%
  - Informed by performance

II. Rule-based Flagging of Errors - Development

- Rules were defined based on **local context** and internal domain knowledge

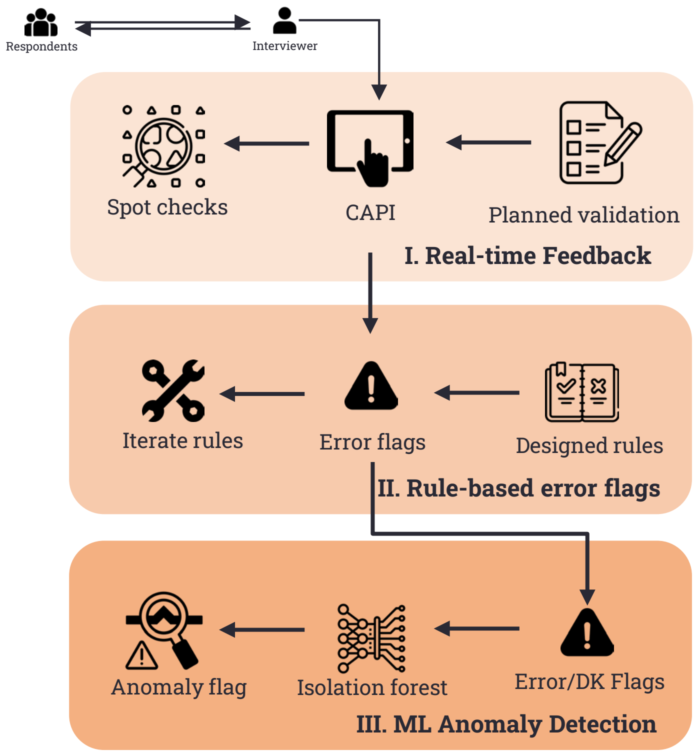
- **New rules were iteratively added** as data was assessed weekly



Contradictory information	Implausible response	Suspicious practices
Respondent can't read a sentence, but reports having a graduate degree	**Husband's age is less than 15 or more than 95**	**Interview completed in under 30 minutes**

III. Anomaly detection using ML

- Used **Isolation forest** - A type of unsupervised Machine Learning Algorithm to identify anomalous records.
  - **Developed by Liu et al (2008)**
  - **Detecting outliers in the data not caught by rules.**
- **Anomalies** are records that differ from the norm in terms of error patterns or frequent "Don't Know" responses – especially where clear answers are expected.



Strengths of Isolation Forest

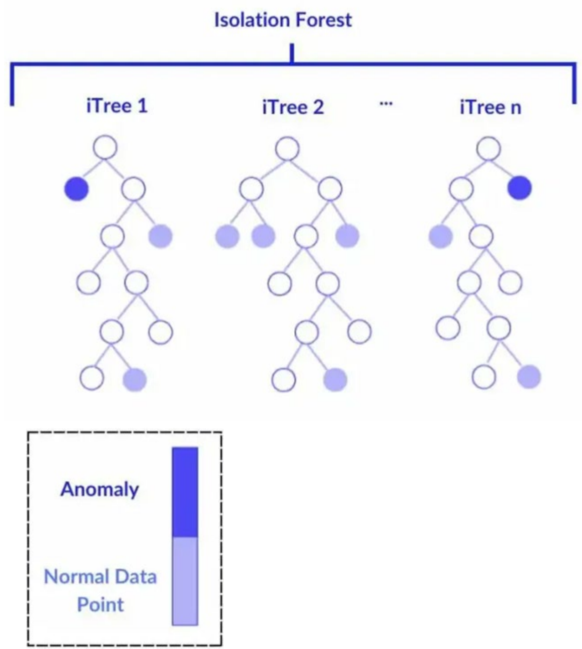


Image credit: https://spotintelligence.com/2024/05/21/isolation-forest/

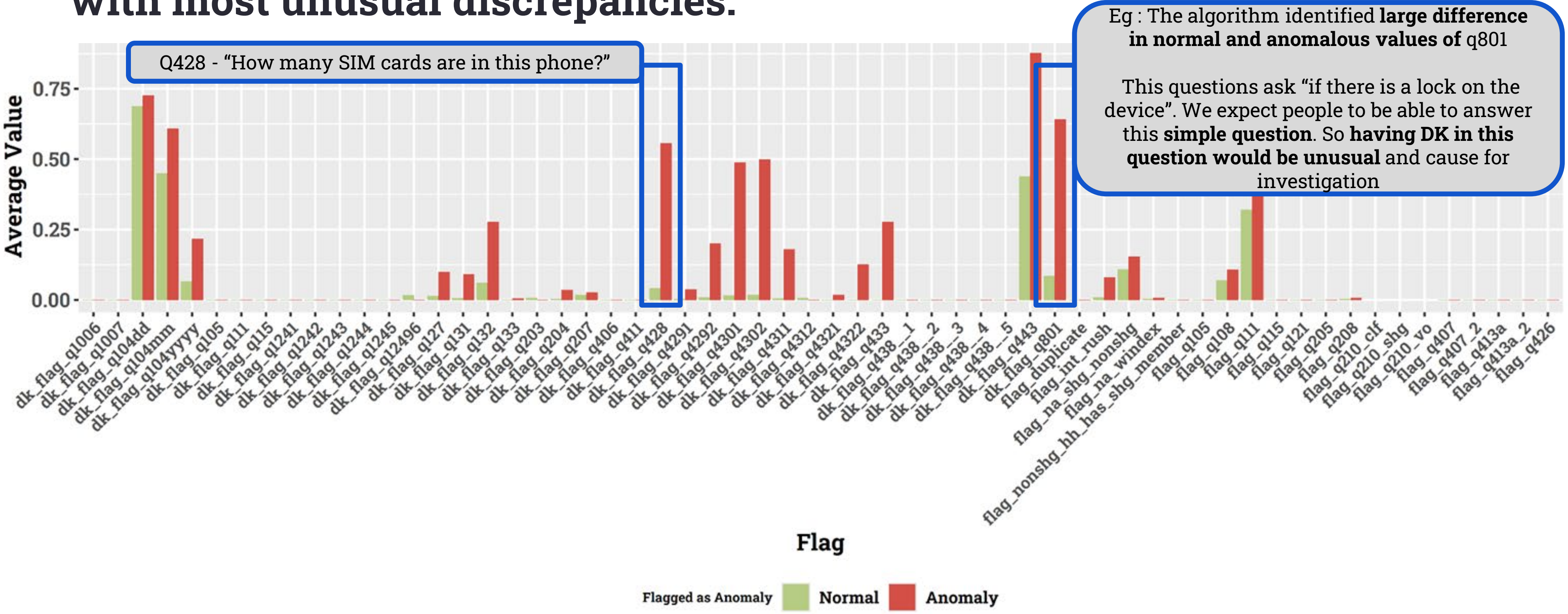
- **Computationally Fast**
- **Generalized.** Does not rely on training data and can accommodate new data easily
- **Low RAM.** Good for large datasets without requiring large computation resource
- Specifically designed for **anomaly detection**
- A 2024 study found it **outperforms* other models** in detection of fabricated records

** Most balanced performance with relatively high precision and recall compared to One-Class Support Vector Machine (SVM), SVM with Stochastic Gradient Descent (SGD), Local Outlier Factor (LOF) algorithm, and Robust Covariance method*

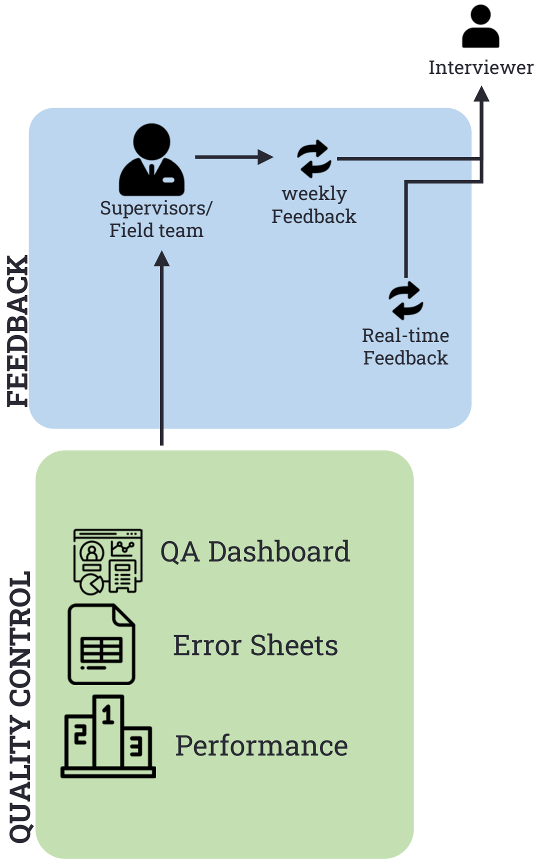
Ref: https://www.sciencedirect.com/science/article/pii/S2468227624003284

How the model detects suspicious records

We assessed variables that showed most **difference** in the average values between normal and anomalous records - ie. **Those variables with most unusual discrepancies.**



Weekly quality control and feedback



Monday	**Data download**	Run QA Scripts -Update dashboard -Generate and share error sheets	Review prior week's data quality
Tuesday	Weekly QA meeting - Dashboard review - Errors discussions - Feedback from field team		**Field team debrief**
Wednesday	Update error flags/rules if required based on feedback and analysis		
Thursday	**Data download**	Run QA Scripts -Update dashboard -Generate and share error sheets	
Friday	**Field team debrief**		

Dashboard Set up

Data Report

Bihar Survey QA Update

PUBLISHED
Saturday Jan 18, 2025 04:16

Sample Tracking

i. Progress Tracking

Tracking completion and response rate of the listing

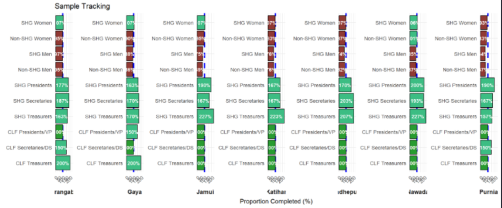
Overall Proportion of SHG Households Listed: 50 %

District	HH_Covered (%)	SHG_HH_Listed (%)	SHG_HH_Covered (%)	HH_Listed
Buxar	0 (0%)	1 (100%)	0 (0%)	1
Aurangabad	601 (9.52%)	3057 (48.42%)	302 (9.88%)	6313
Gaya	575 (8.51%)	3408 (50.41%)	295 (8.66%)	6760
Katihar	[illegible]	[illegible]	[illegible]	[illegible]

ii. Individual Interview Tracking

No. of valid interviews= ifelse(consent == 1 & interview__status == 100, 1, 0)

District	Completed SHG Women	Target SHG Women	Completed Non-SHG Women	Target Non-SHG Women	Completed SHG Men	Target SHG Men
Buxar		375		375		300
Aurangabad	403	375	357	375	291	300
Gaya	400	375	338	375	263	300
Katihar	364	375	347	375	228	300



Quality Analysis

a. Cumulative Errors Summary

Search:

enm_name	sup_name	total_cases	total_flags	flag_duplicate	flag_int_rush	f
Amit Goswami	Santanu Maji	119	49	0 (0%)	10 (8.4%)	0
Prashant Kumar	Jitendra Meena	159	45	0 (0%)	6 (3.8%)	0
Saumya Patel	Rajesh Kumar Singh	115	45	0 (0%)	8 (7%)	0
[illegible]	Sushil	[illegible]	[illegible]	[illegible]	[illegible]	

Showing 1 to 264 of 264 entries

b. Cumulative Errors Sheet

Search:

enm_name	sup_name	total_flags	flag	link_id	data_source
Amit Goswami	Santanu Maji	49	flag_int_rush	2_18_241_162_1_0	wom_data
Amit	Santanu	49	flag_q108	2_18_241_162_1_0	wom_data

c. Weekly Error Summary

Week: 7 (18 Jan - 24 Jan)

Total flags this Week: 79

Flag Counts

Search:

flag_name	count
flag_anomaly	31
flag_na_windex	15
flag_q407	12
flag_q413a	7
flag_q208	3

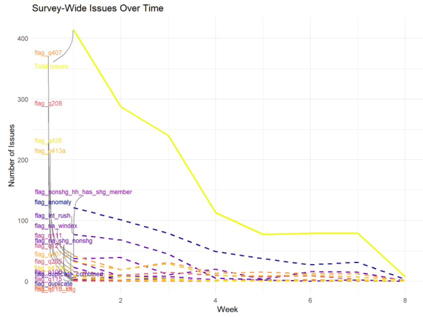
Showing 1 to 10 of 10 entries

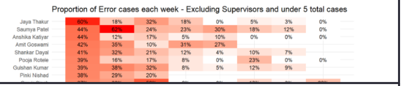
Weekly Errors Summary for each Enumerator

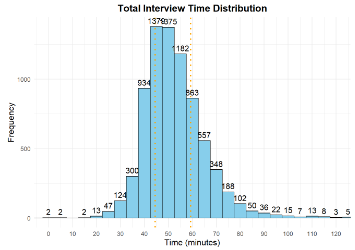
Search:

enm_name	sup_name	total_cases	total_flags	flag_duplicate	flag_int_rush	f
Ghanshyam Sahu	Bibekananda Prusti	4	3	0 (0%)	0 (0%)	0
Gulshan Kumar	Sushil Kumar	23	3	0 (0%)	0 (0%)	0

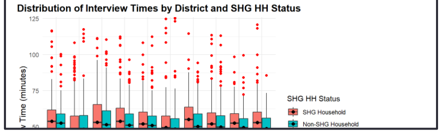
d. Enumerator Performance








```
clip_dis_time_box <- dis_time_box +
  coord_cartesian(ylim = c(0, 120))
clip_dis_time_box
```



Weekly Error Sheets

- Field team used this to investigate and correct errors on a case by case basis.
- Example: Week 5 Error Sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	enm_nam	sup_name	total_flag	flag	ID	data_source	interview	total_dk_flag	TotalNA	week	q105	q111	q115	q121	q407	q413a	q205	q207	q208
2	SP	RK	9	flag_q407	5_66_355_89_2	wom_data	04-80-05-77	2	603	5	37	17	40	20	1				
3	SP	RK	9	flag_anomaly	5_66_355_89_2	wom_data	04-80-05-77	2	603	5	37	17	40	20					
4	SP	RK	9	flag_anomaly	5_71_367_180_	wom_data	07-04-88-90	5	540	5	35	18	36	19			12	2	2
5	SP	RK	9	flag_anomaly	5_71_367_181_	wom_data	12-49-50-00	6	585	5	28	17	32	21	24				
6	SP	RK	9	flag_anomaly	5_71_367_46_1	wom_data	17-39-14-68	5	574	5	50	18	51	19					
7	SP	RK	9	flag_anomaly	5_71_367_138_	wom_data	32-77-76-91	6	575	5	52	17	53	19			12		
8	SP	RK	9	flag_q407	5_71_367_10_1	wom_data	35-77-12-17	3	561	5	28	18	30	22	3		15		
9	SP	RK	9	flag_na_winde	5_66_355_172_	wom_data	47-77-96-60	4	562	5	36	19	38	23	34				
10	SP	RK	9	flag_anomaly	5_66_355_152_	wom_data	64-65-41-96	3	558	5	55	17	60	19	53		12		
11	AG	SM	7	flag_anomaly	5_60_342_161_	wom_data	44-09-73-43	4	535	5	36	18	40	25		30	12		
12	AG	SM	7	flag_anomaly	5_67_357_70_3	wom_data	92-73-57-37	4	526	5	38	18	45	20	30		10		
13	AG	SM	7	flag_anomaly	5_67_357_70_4	shg_data	34-09-59-35	6	536	5	56	15	65	17		45	11		
14	AG	SM	7	flag_anomaly	3_103_601_75_	shg_data	72-61-81-65	7	501	5	48	16	55	18		40	10		
15	AG	SM	7	flag_anomaly	5_67_357_195_	men_data	28-09-60-66	6	504	5	45	20	40	22		35			
16	AG	SM	7	flag_anomaly	3_103_601_58_	men_data	67-20-21-15	5	455	5	51	20	48	22		40			
17	AG	SM	7	flag_anomaly	3_103_601_78_	men_data	98-14-40-39	6	523	5	60			17		50			
18	SL	BP	6	flag_q208	5_69_364_111_	wom_data	01-05-47-52	3	503	5	32	16	35	17	30		12	3	1
19	SL	BP	6	flag_q413a	3_102_470_224	wom_data	14-37-98-29	3	522	5	22	18	25			1			
20	SL	BP	6	flag_q413a	5_63_349_19_3	wom_data	47-81-30-77	3	464	5	24	18	26			1			
21	SL	BP	6	flag_q413a	5_63_349_67_1	wom_data	51-39-68-64	3	526	5	41	15	45	18		1			
22	SL	BP	6	flag_q208	5_63_349_130_	wom_data	67-70-19-21	4	389	5	20	16	23	17	16		12	98	4
23	SL	BP	6	flag_q208	5_69_364_108_	shg_data	12-05-72-45	3	526	5	34	18	36	19		22	12	2	1

Age of first mobile phone

Q207: Reported last SHG meeting took place more than 1 month ago

Q208: Reports attended SHG meeting a week ago

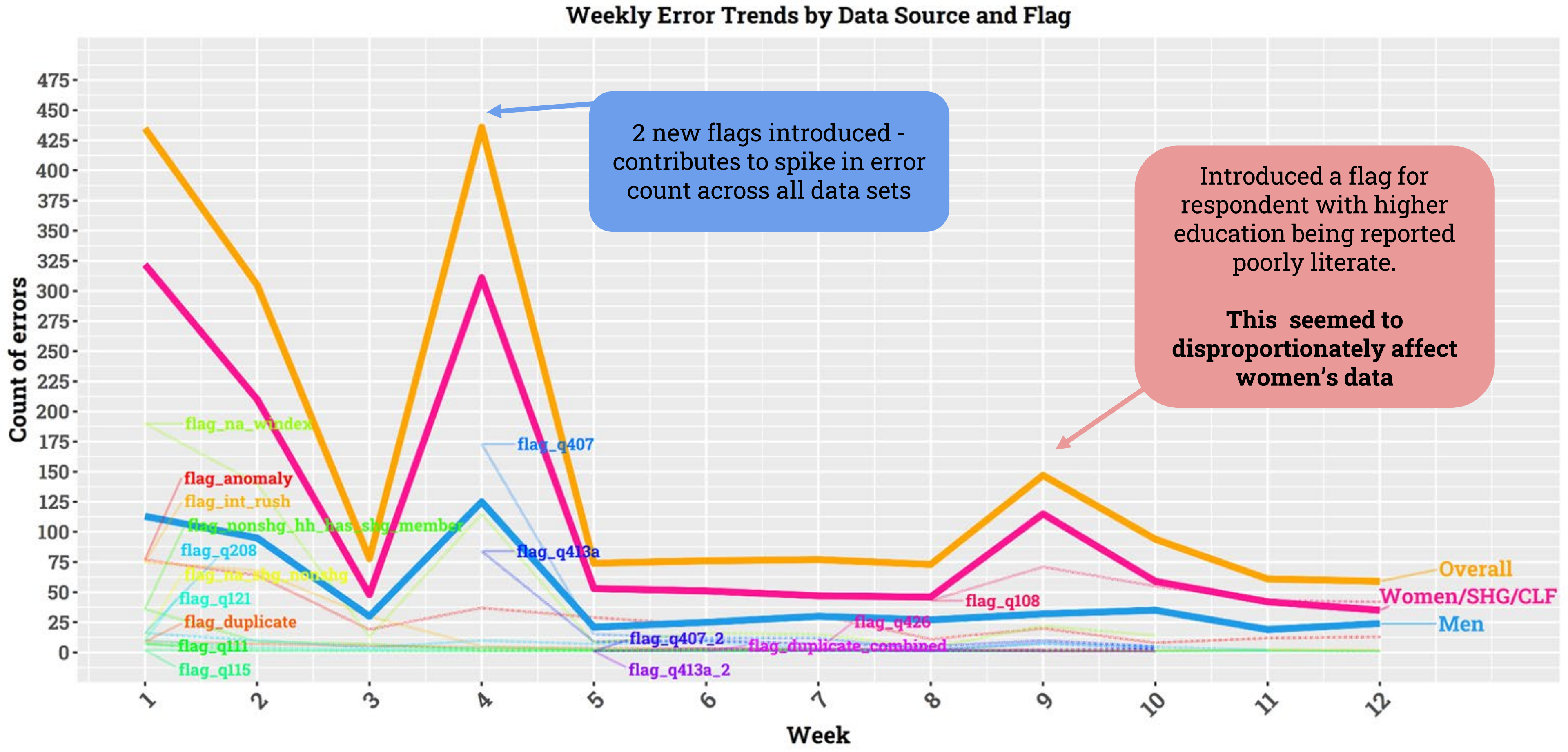
Enm and Sup names

Error type

Interview identifiers

Compare values in records

85% improvement in error rates over 12 weeks



Key Performance Indicators

- All sample size and data collection targets were met within scheduled time frame
- Observed **85% drop in error rates** across the course of data collection
- Observed **20 - 50 % point decrease in error rates** for interviewers with higher error rates at start of data collection

Next steps

- **Integrate paradata -** Keystroke tracking, timestamps, GPS tracking, and audio capture **where possible**.
- **AI/LLM integration**. Explore use of large language models for real-time curbstoning detection, ensuring privacy safeguards.
- **Open-source resources.** Share rules and anomaly detection pipelines to enable replication and scaling.

Anticipated use for QA / QC approach

- **Improve survey data quality**
- **Opportunity to apply this approach to improve data quality more widely to a range of development programs**
  - End to end digitization of health systems
  - Government efforts in India to digitize Self-help groups for economic empowerment

Questions?

Thank you

on behalf of the EDiT Team

Dr. Mayank Date

mdate1@jhu.edu

Data Scientist, Johns Hopkins University

Project funded by
Gates Foundation

evidence-digital.org

Scan to learn more about
Evidence for Digital Transformation (EDiT)