IPSI

Module 4.1
Right Data – Right Decisions: Approach to Data Quality

Trainer:.....

Module objectives

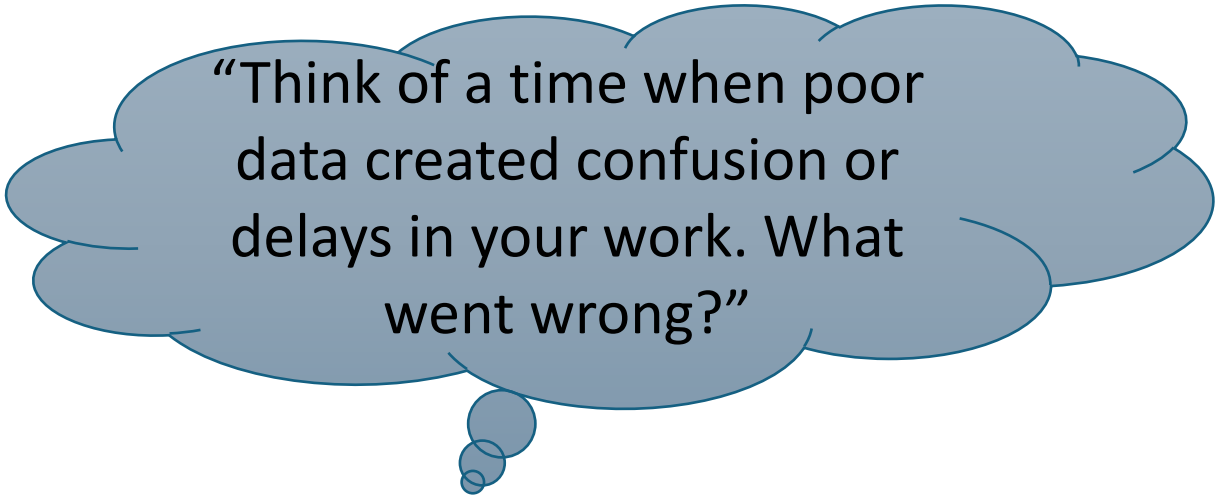
At the end of this module, participants will be able to:

1. Identify the types of data errors that affect the quality of health data.
2. Define data quality and identify the key dimensions of data quality assessment.
3. Apply techniques to improve the quality of health data.

Module Outline

S. No	Session name	Session format	Time
1	Data quality and its importance	Facilitator led presentation	20 min
2	Data quality assessment and methodology- a) Completeness b) Inconsistency (Outliers) c) Inconsistency	Facilitator led presentation and exercise	120 min

Why data quality matters and why should we care?



Data Informs. Poor Data Misleads.

Data quality is about whether the data we collect and use is fit for its intended purpose — whether it can be trusted and used with confidence!

- Planning, monitoring, reporting, and evaluation all depend on one thing — quality data.
- In public health, data isn’t just numbers — it’s the foundation of every good decision: poor data, ineffective programs.

"*Data quality is a critical component of effective public health systems, and errors can compromise data usability and decision-making*.” (McGuckin et al., 2022; Maïga et al., 2019)

Poll questions

1. Why do you care about data quality?
   a) Because I like making smart decisions
   b) Because my supervisor does 😅
   c) Because poor quality data once made me look bad 🙈
   d) Wait... am I supposed to care?

2. Has poor quality data ever messed up your work?
   a) Oh yes, big time!
   b) Just once—lesson learned
   c) Not that I know of (fingers crossed)
   d) What even is ‘poor quality’ data?

3. What happened when poor data showed up in your work?
   a) Got scolded in front of everyone 😬
   b) Quietly fixed it and pretended it never happened 😎
   c) Blamed Excel/ portal/ application/ software and moved on
   d) Surprisingly... I got praised for spotting it!

Data Quality Issues in Health Systems Data

Data quality issues	Health systems examples
1. Missing data	Blank fields in registers, delayed reports, non-reporting by health facilities by the deadline
2. Inconsistency (across sources/time)	Reports show higher ANC coverage than survey data; year-to-year trends fluctuate unusually
3. Measurement & entry errors	Wrong BP recorded, age misclassified, transcription/typing mistakes during data entry
4. Issues in definitions & indicators	Different understanding of “fully immunized child”; unclear reporting formats
5. Manipulation & bias	Inflated reporting to meet targets; under-reporting to hide poor performance

Data Quality Issue – Missing Data

Reasons	Consequence	Methods to identify errors
• **Blank fields** in registers • **Non-reporting** by facilities • **Partial reporting** – some indicators submitted, others missing	• Under-reporting (numerators) • Unreliable denominators • Data gaps	• Completeness checks (count blanks, compare expected Vs received reports) • Automated dashboards • Annual desk review

Potential sources of missing data

- Data was not recorded or reported; Services were not available in the facility.

Data Quality Issue – Missing Data

Reasons	Consequence	Methods to identify errors
• **Delayed reports** – late submission of monthly/ quarterly forms • **Irregular reporting** – inconsistent cycles across facilities/districts	• Data not available for real-time decision-making • Delays in interventions	• Timeliness tracking (submission logs) • Timeliness indicators

Metrics for Assessing Missing Data & Delayed Reports

Completeness of data

- Measures whether all expected entities (facilities/blocks) submit reports
- All data elements/ points reported
- *Applies to both facility-to-block and block-to-district reporting*

Timeliness of Data

- Measures if reports are submitted before the set deadline

Completeness and Timeliness Indicative Metrics

Data quality metric	Definition with examples
District Reporting Completeness	Benchmark: ≥ 90% of expected monthly reports received. *Poor performance*: < 75% completeness.
Facility Reporting Completeness	Benchmark: ≥ 80% of facilities submitting all expected reports. *Poor performance*: < 75% of facilities reporting.
Indicator Data Completeness (Non-missing/ Non-zero values):	**Maternal Health (e.g., ANC1)**: ≥ 90% **Malaria/HIV**: ≥ 90%
District Timeliness	Benchmark: ≥ 75% of submitted reports received on time. *Poor performance*: < 75% timeliness.

Source: WHO Data Quality Review (DQR) Toolkit

Completeness: Data Quality metrics

Data quality metric	Examples
Percentage of units with submitted reports that have non-missing values for the selected indicator(s) within the reporting period month.	ANC 4 visits: % of facilities that reported data on ANC 4 visits in a month.
Percentage of units with missing values entered for the selected indicator(s) within the reporting period	Missing Values: % of facilities with more than 10% missing values for the chosen indicator(s) or data point. (10% is just indicative, not a benchmark or threshold)

Operational Methods to Address Missing data Issues

- **Regular training and supportive supervision**
  - ASHAs, MPHWs/ANMs, CHOs, Data entry operators
- **Automated Alerts/Reminders**
  - SMS/WhatsApp/email prompts to facilities that miss reporting deadlines.
- **Supervisory Feedback Loops**
  - Monthly review meetings where missing reports are highlighted and followed up.
- **Peer Benchmarking**
  - Rank facilities on completeness; use positive competition.

Group exercise 1

- Identify the missing values and zeros reported in ANC services data
- Calculate the percentage of zeros and missing values and interpret the results.

Data Quality Issue – Inconsistency

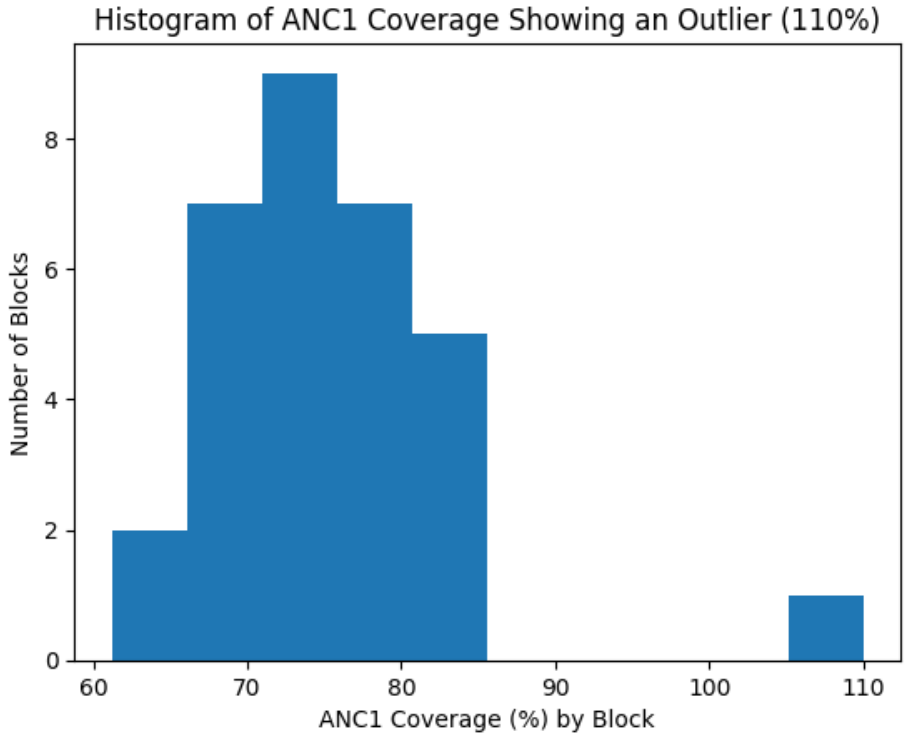
Data quality issue	Consequence	Methods to Identify Errors
• **Presence of outliers** – presence *of impossible/ extreme values* • **Inconsistency over time** – unrealistic fluctuations in indicators Q-Q or Y-Y • **Inconsistency between sources** – HMIS vs. DHS/MICS/survey data • **Internal inconsistency** – mismatch between related indicators (e.g., ANC1 vs. TT1)	• Lack of comparability • Confusion • Reduced trust in data	• Desk review of trends • Outlier detection • Ratio checks between related indicators • Triangulation with surveys

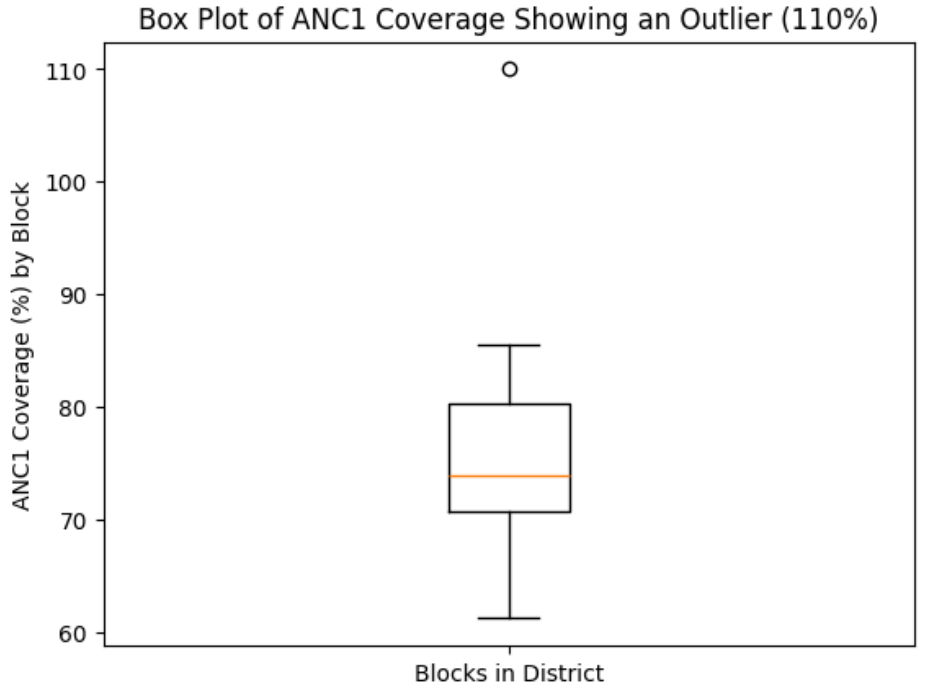
Inconsistency 1 - Outliers

Data quality metric	
Definition	Outliers are numbers that do not conform to the trend or are unexpected (extreme) values.
Indicative Metric from WHO	*Acceptable*: Fewer than 5% of values flagged as moderate outliers; extreme outliers should be rare.
Eyeballing (Visual Check):	Simply look at the data series or graph. Sudden spikes or dips compared to other months often stand out.
SD method	**Histogram**
IQR method	**Outliers may be represented using a box plot.**

Inconsistency 1 - Outliers

- Monthly district review meeting: ANC1 coverage was reported across 30 blocks in the district, between 65% and 85%, consistent with previous months and expected programme performance. However, one block reports ANC1 coverage of 110%.





Outlier detection using SD

- In statistical terms, outliers are the values lies beyond 1.5 SD in the range 2 or 3 SD can be taken based on context.
- In formula terms - Upper limit is Mean + 1.5 × SD, and Lower limit = Mean − 1.5 × SD
- This often helps to identify data entry errors or large deviations in computation.
- Let's take previous example of OPD per 1000 population in a block
  - ➢ Facility level data points that are more than 1.5 SD from the mean.

Block A Number of SHCs = 23		
Monthly OPD	**Population**	**OPD per 1000**
183	9,836	18.64
305	7,122	42.78
316	9,539	33.09
341	4,315	78.95
134	7,101	18.87
213	4,795	44.42
251	11,040	22.77
291	9,344	31.14
123	3,676	33.46
283	10,478	26.98
399	7,861	50.71
249	8,900	27.94
161	5,165	31.24
68	1,050	64.76
257	5,046	51.00
80	6,194	12.97
234	5,684	41.17
242	1,312	184.71
272	6,806	40.01
287	6,912	41.57
260	6,790	38.29
73	4,263	17.20
218	7,934	27.43

- Mean OPD per 1000 population= 42.6
- Standard deviation= 34.6

Outliers are values that are greater than mean by (1.5*34.6)

-> 42.6 + (1.5*34.6)= 94

Values greater than 94:
One value= 184.7

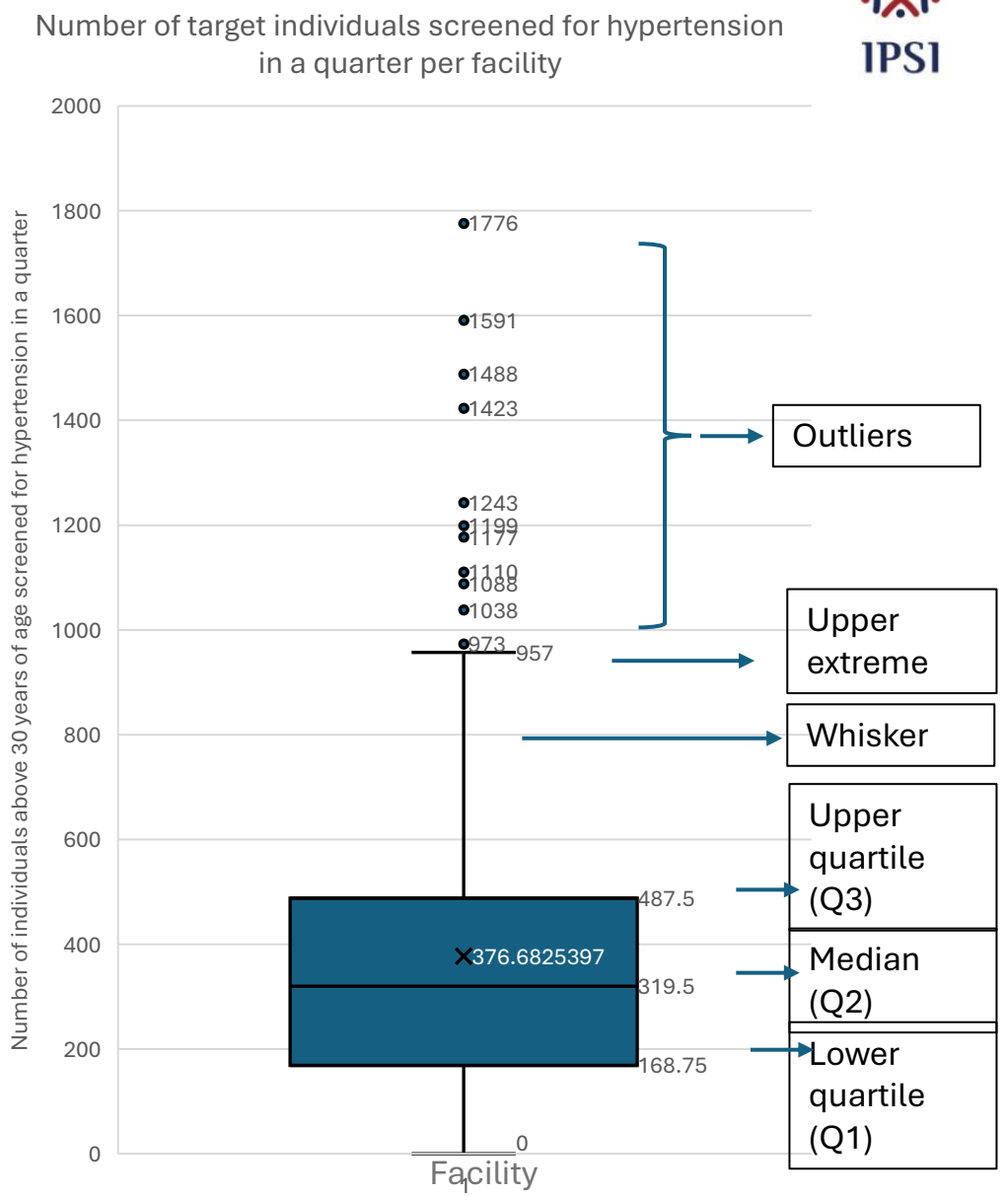
Outlier detection using box plot Example from previous module

Box Plot:

- Shows Q1, Q2, Q3 values clearly.
- Displays Interquartile Range (IQR = Q3 – Q1) → spread of middle 50% of data.
- Identifies outliers, values
  - < Q1 – 1.5×IQR *OR*
  - > Q3 + 1.5×IQR

Identifying Outliers

- Outliers = Data points beyond 1.5×IQR from Q1 or Q3.
- Represented as dots, stars, or asterisks in the plot.



Data source: NCD Portal, X district

Operational Methods to Address Outliers

- **Verification of outliers**
  - Contact the facility or verify with data sources, such as registers or special drives.
- **Rectification of outliers**
  - Data entry issue - correction
  - Genuine outliers - retain
  - Unverified outlier - flagged till verification
  - Keeping log and sharing feedback
- **Preventive measures**
  - Training of those involved in data entry
  - Reinforce indicator definitions/ protocols to avoid misreporting
  - Suitable validation rules among portals
- **Recognize facilities with “nil” outliers**

Group exercise 2

- Identify the outlier facilities for the following data:
  1. ANC registration- using SD method
  2. TB screening- using box plots
- Interpret the results

Inconsistency 2 - Inconsistency over time

Data quality metric	
Consistency over time	Checks whether reported values follow a logical/ expected trend across time (for each identified indicator).
Recommended rule-of-thumb value	Current year values should be within ±33% of the average of the previous 3 years (for stable indicators). (The threshold of 33% is indicative; the relevant level may be decided by program managers in the given context for analysis purposes.) For increasing/decreasing trends, compare to forecasted values.
Metric example	Number and percentage of facilities (or blocks) where the current month's value for the selected indicator differs by more than ±33% from the **district average** over 1/2/3 years Number and percentage of facilities (or blocks) where, for any of the selected tracer indicators, the current month's value differs by more than ±33% from the **district average**.

Inconsistency over time

Total No of ANCs over time

Facility	Feb-25	Mar-25	Apr-25	May-25	Jun-25	Jul-25	6m Avg	Aug-25	Facility ratio	District Ratio	Difference	Deviation	Flag
Facility A	100	110	95	105	90	100	100	110	1.10	**0.89**	0.21	24%	No
Facility B	82	75	70	92	80	91	82	40	0.49	**0.89**	-0.40	-45%	Yes
Facility C	90	96	88	92	78	90	89	90	1.01	**0.89**	0.12	14%	No
Total for district	272	281	253	289	248	281	271	240	**0.89**	**0.89**			
								Current Month	**= Current mo/ 6m Avg**	**= Current mo/ 6m Avg**	**= Facility – District ratio**	**= Difference / District ratio**	

Inconsistency 3 - Inconsistency between related indicators

Data quality metric	
Definition	Examines expected relationships between related data points/indicators
Indicative Metric	For indicators which should be ≥ 1, districts with ratios of < 1 should be flagged. The ratio percentage difference between state and national ratios is calculated. District units with a percentage difference greater than the specified quality threshold (e.g. ≥ 10%) are flagged for follow-up.
Example metric	1. Consistency between indicators (within the same facility, across related data points, where there is a difference (± 10%) between the indicator) Examples ▪ Maternal health: ANC1 = TT1 ▪ Immunization: Pentavalent dropout rate, (PV1 – PV3/PV1) – should not be negative ▪ NCD: No of cases diagnosed for hypertension should be equal to the number of cases initiated on treatment, at the facility level 2. Consistency across facilities (compare *how facilities differ from each other (or the district average)* for the same indicator relationship) Number of facilities where there is a difference (± 10%) for each indicator
Method	▪ Select logically related core indicators. ▪ Compare the facility-level ratio with the district ratio. ▪ Set a threshold for acceptable deviation. ▪ Flag facilities that exceed the threshold for review

Consistency between related indicators

Total No of ANC1 Vs TT1

- Related indicators: ANC1 ≈ TT1
- Method:
  - Compare ANC1 and TT1 at the facility level
  - Allowable deviation: ±10%
  - Flag facilities exceeding threshold
- Example (District X, May 2025):
  - Facility A: ANC1=100, TT1=98 → Consistent
  - Facility B: ANC1=120, TT1=90 → Flag
  - Facility C: ANC1=75, TT1=74 → Consistent
- Metric: % of facilities with ≥10% difference

Inconsistency 4 - Inconsistency with original records

Data quality metric	
Definition	Verification of reporting consistency through facility survey
Metric	% agreement between verified counts for selected indicators in sampled facility records and reported values for the same facilities
Method	• A sample of health facilities will be selected for the data verification exercise. • Original source documents for selected indicators will be reviewed at these facilities. • Reported data will be compared against the records in the source documents. • Consistency will be measured by calculating the agreement between the reported and source data. • This exercise involves primary data collection through record reviews and is the only internal consistency dimension assessed this way.

Inconsistency 5 - External comparison and inconsistency

Data quality metric	
Definition	Comparison of routine data with population-based survey values from the same period
Method	Comparisons can be made: Program Data ↔ NFHS (Survey Data) Program Data ↔ HMIS (Facility Reporting Data) Indicators may be compared either • Through ratio (Program data ÷ Survey data) **or** • By observing the difference between values.
Indicative Metric	Routine vs. survey coverage rates should be within ±33%.
Example	Indicator: Mothers with ≥4 ANC Visits (ANC Coverage) **Data Source** — **ANC1 Coverage** a) RMNCH+A (Program data) — 92% b) HMIS data — 89% c) NFHS — 82% **Comparison and interpretation:** (i) Ratio (Program data ÷ NFHS data) = 92 ÷ 82 = 1.12 → Program data reports 12% higher coverage than NFHS, suggesting possible over-reporting (ii) Program vs. HMIS difference = 92% vs. 89% → Shows small variation, but possible reporting gaps.

Operational Methods to address inconsistency over time issues

- **Cross-verification with health facility records**
  - Contact the facility or verify with data sources, such as registers or special drives
- **Correction**
  - If the error is verified, the unverified outlier – flagged till verification
  - Identify the reason and documentation, feedback sharing
  - Keeping log and sharing feedback
- **Preventive measures**
  - Training of those involved in data entry
  - Reinforce indicator definitions/ protocols to avoid misreporting
  - Suitable validation rules among portals
  - Recognize facilities with “nil” outliers

Group exercise 3

- Identify the inconsistency between related indicators
  - OPV-0 (birth dose) and Hepatitis-B0 (birth dose)
- Identify the inconsistency across time
  - Facility-wise monthly data for approximately four months on new ANC registrations.
- Interpret the results

Data quality issues related to Accuracy

Issue	Consequence	Methods to Identify Errors
• **Measurement & entry errors** – wrong BP recorded, duplicates, transcription/coding mistakes, tallying errors • **Issues in definitions & indicators** – varied understanding of "fully immunized child," ambiguous reporting forms • **Manipulation & bias** – inflated numbers to meet targets, under-reporting failures	• Faulty analysis, • incorrect programmatic action, • Loss of credibility and integrity	• Facility data verification (registers vs. reports) • Audits • Supervisory reviews • Indicator standardization

Measurement & entry error examples

- **Accuracy** refers to the correctness of data reported

Data entry errors			
Indicator	Correct entry	Typing error	Wrong field entry
ANC registration	**100**		**80**
Registration in 1st trimester	**80**	**805**	**100**

Issues in definitions & indicators example

Question: To calculate the total number of children fully immunized

Child ID	BCG	Penta1	Penta2	Penta3	OPV0	OPV1	OPV2	OPV3	Measles	Fully Immunized (Correct)
Child_1	1	1	1	1	1	1	1	1	1	1
Child_2	1	1	1	1	1	1	1	1	1	1
Child_3	1	1	1	1	1	1	1	1	0	0
Child_4	1	1	1	1	1	1	1	1	1	1
Child_5	1	1	1	0	1	1	1	1	1	0
Child_6	0	1	1	1	1	1	1	1	1	0
Child_7	1	1	1	1	1	1	1	1	1	1
Child_8	1	1	0	1	1	1	1	1	1	0
Child_9	1	1	1	1	1	0	1	1	1	0
Child_10	1	1	1	1	0	1	1	1	1	0
										4

Child ID	BCG	Penta1	Penta2	Penta3	OPV0	OPV1	OPV2	OPV3	Measles	Fully Immunized (Incorrect)
Child_1	1	1	1	1	1	1	1	1	1	1
Child_2	1	1	1	1	1	1	1	1	1	1
Child_3	1	1	1	1	1	1	1	1	0	0
Child_4	1	1	1	1	1	1	1	1	1	1
Child_5	1	1	1	0	1	1	1	1	1	0
Child_6	0	1	1	1	1	1	1	1	1	0
Child_7	1	1	1	1	1	1	1	1	1	1
Child_8	1	1	0	1	1	1	1	1	1	0
Child_9	1	1	1	1	1	0	1	1	1	0
Child_10	1	1	1	1	0	1	1	1	1	0
	9	10	9	9	9	9	10	10	9	84

✅ **Correct compilation**

Only those children who have received BCG, all doses of Pentavalent and OPV, and have received the Measles dose during the month will be counted as fully immunized.

❌ **Incorrect compilation**

Adding all children and reporting them as fully immunized

Accuracy: indicative metric from WHO

Data quality metric	Definition with examples
Measurement and entry error	Reported values should be within **±10%** of values verified through source documents.

Source: WHO Data Quality Review (DQR) Toolkit

Operational Methods to Address Accuracy Issues

- **Standardize protocol**
  - ➢ Clear definition of indicators, clear step-by-step guide for measurement as well as data ranges
- **Training**
  - ➢ Provide routine training for portal/ users to ensure full understanding of correct ways of inputting data
- **Data quality monitoring and reporting**
  - ➢ Proactively seek potential accuracy issues when seeing them

Source: https://www.atlassystems.com/blog/data-quality-in-healthcare

Cultivating a culture of data quality through structured communication

- **Situation, Background, Assessment, Recommendation (SBAR) Framework**
  - Used in clinical communication, can be adapted to improve public health data quality.
  - Promotes a shared, structured, and actionable way to identify and resolve data issues.
  - Encourages quality at every stage of the data cycle (data entry to reporting to feedback)
- **Key Benefits of Using SBAR**
  - Promotes ownership and accountability
  - Enables clear, standardized communication
  - Supports continuous improvement in data practices

Applying SBAR to a Real-Life Data Problem

SBAR Component	Meaning	Application: Incorrect computation of fully immunized children
S – Situation	Describe the current problem or abnormality detected.	We identified that the reported number of fully immunized children seems abnormally high.
B – Background	Provide relevant context, history, or process details that led to the situation.	The same person isn't compiling and entering the data. The method used was summing across children, and instead of verification by each child
A – Assessment	Analyze the problem, identify likely causes, and its potential impact.	This is likely due to gaps in understanding the indicator definition, or new staff may have joined or lack of communication between providers
R – Recommendation	Suggest specific actions to address and prevent recurrence.	Conduct training on immunization indicator computation. Develop a checklist for compiling monthly reports correctly for ease of calculation by those who compile

Poor Data Quality - Real-World Impact

Poor data quality

- ✓ Errors in collection, entry, analysis or reporting
- ✓ Incomplete, inconsistent, inaccurate data

System consequences

- ✓ Misallocated resources
- ✓ Misleading performance rankings
- ✓ Weak planning & monitoring
- ✓ Compromised patient safety and public health decisions
- ✓ Loss of trust

District Impact

- ✓ Inefficient programs
- ✓ Missed health targets
- ✓ Lower community confidence
- ✓ Poor health outcomes

References

1. Toolkit for analysis and use of routine health facility data: general principles. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.

2. Data quality review: a toolkit for facility data quality assessment. Module 1. Framework and metrics. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.

3. Data quality review: a toolkit for facility data quality assessment. Module 2. Desk review of data quality. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.

4. Toolkit for analysis and use of routine health facility data: integrated health services analysis: district and facility level. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.

5. NHSRC website. Available from: https://nhsrcindia.org/sites/default/files/2021-04/Session%203%20Presentation%20for%20Volume%202%20Health%20Programme%20Managers%20Manual%20Data%20Quality_0.pdf

6. Swerdlow, A.J. (2014). Data Quality in Vital and Health Statistics†. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat06088

References

7. McGuckin T, Crick K, Myroniuk TW, Setchell B, Yeung RO, Campbell-Scherer D. Understanding challenges of using routinely collected health data to address clinical care gaps: a case study in Alberta, Canada. BMJ Open Quality [Internet]. 2022 Jan 1;11(1):e001491. Available from: https://bmjopenquality.bmj.com/content/bmjqir/11/1/e001491.full.pdf

8. Maïga A, Jiwani SS, Mutua MK, Porth T, Taylor CM, Asiki G, et al. Generating statistics from health facility data: the state of routine health information systems in Eastern and Southern Africa. BMJ Global Health [Internet]. 2019 Sep 1;4(5). Available from: https://gh.bmj.com/content/4/5/e001849

9. Chen H, Yu P, Hailey D, Wang N. Methods for assessing the quality of data in public health information systems: A critical review. Studies in health technology and informatics [Internet]. 2014 Jan 1;204:13–8. Available from: https://pubmed.ncbi.nlm.nih.gov/25087521/

10. Johnson, S. C., Cunningham, M., Dippenaar, I. N., Sharara, F., Wool, E. E., Agesa, K. M., Han, C., Miller-Petrie, M. K., Wilson, S., Fuller, J. E., Balassyano, S., Bertolacci, G. J., Weaver, N. D., Lopez, A. D., Lopez, A. D., Lopez, A. D., Murray, C. J. L., Murray, C. J. L., Naghavi, M., & Naghavi, M. (2021). Public health utility of cause of death data: applying empirical algorithms to improve data quality. *BMC Medical Informatics and Decision Making*, *21*(1), 175. https://doi.org/10.1186/S12911-021-01501-1

11. Naghavi M, Richards N, Chowdhury H, Eynstone-Hinkins J, França EB, Hegnauer M, et al. Improving the quality of cause of death data for public health policy: are all “garbage” codes equally problematic? BMC Medicine [Internet]. 2020 Mar 9;18(1):1–3. Available from: https://link.springer.com/content/pdf/10.1186/s12916-020-01525-w.pdf

Key takeaways: Assessing data quality

- The quality of health data is primarily hindered by missing data, outliers, inconsistent and inaccurate data.
- Key dimensions of data quality assessment are:
  -Completeness, timeliness
  -Consistency (outliers, consistency across related indicators, consistency across time)
  -External validity
  -Accuracy
- There are a wide variety of techniques that can be used to prevent each type of data quality issues or improve the quality of existing data.

Thank You