

Four Practical Indicators for Real-Time Verbal Autopsy Data Quality Assessment: A Cross-National Validation Study

Isaac Lyatuu

ilyatuu@ihi.or.tz

Ifakara Health Institute

Mahadia Tunga

University of Dar-es-Salaam

Shukurani Irema

Ifakara Health Institute

Gisbert Msigwa

Ministry of Health

Trust Nyondo

Ministry of Health

Claud Kumalija

Ministry of Health

Makhosazana Mokoena

Ministry of Health, Kingdom of Eswatini

Bukar Zarami

Clinton Health Access Initiative (CHAI)

Shamsuddeen Yahaya

Katsina State Primary Health care Agency

Robert Mswia

Vital Strategies

Research Article

Keywords: verbal autopsy, data quality, completeness, reliability, internal consistency, interview duration, WHO-VA instrument, mortality surveillance

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10907895/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Verbal autopsy (VA) data quality is conventionally assessed after or during cause-of-death (CoD) assignment, a step that requires aggregated data and introduces delays in identifying and correcting data problems. While field training and electronic form logic (e.g., skip patterns, range constraints) address basic data collection quality at the point of entry, they do not systematically evaluate respondent engagement, respondent reliability, logical consistency across clinically related fields, or interview process efficiency. To our knowledge, no standardized, instrument-agnostic set of quality indicators has been published that can be applied to individual VA records in real time, prior to or alongside CoD assignment, to capture dimensions of data quality that entry-level validation does not address. We propose and validate four complementary indicators: the Informative Completeness Score (ICS), Respondent Reliability Score (RRS), Internal Consistency Index (ICI), and Median Interview Duration (MID), that can be computed for a single VA record, enabling immediate feedback to interviewers and supervisors distinct from and complementary to field-level validation.

Methods

The four indicators were developed using field names and response codes defined in the 2016 and 2022 World Health Organization Verbal Autopsy (WHO-VA) instruments. We validated the indicators on three countries' datasets: Tanzania (n = 4,344, 2016 WHO-VA), Eswatini (n = 380, 2022 WHO-VA), and Nigeria (n = 5,474, 2022 WHO-VA). Descriptive statistics, tier distributions, and per-rule violation frequencies were computed for each indicator.

Results

ICS was high across all datasets (means: Eswatini 96.4%, Tanzania 97.4%, Nigeria 99.8%), with occasional low-outlier records indicating interviewers who relied heavily on "don't know" or "refused" responses. RRS was computable for Eswatini (n = 349, mean = 81.8, 63.3% High tier) and Tanzania (n = 3,861, mean = 84.1, 67.1% High tier); the Nigeria dataset lacked the required relationship and date fields. Across nine ICI consistency rules, overall performance was Excellent in all three datasets (Eswatini 91.8%, Tanzania 92.8%, Nigeria 92.0%). Cough-duration violations (C4) were most frequent in Eswatini (6.1% vs 1.6% in Tanzania); Nigeria showed the highest number of fever-duration violations (C3, 2.1%) alongside a markedly elevated rate of pregnancy-related symptoms reported for male decedents (C7, 5.9%); and Tanzania showed an elevated rate of interviews dated before the recorded death (C9, 3.5%), consistent with a discrete period of reduced data-entry oversight. MID was computable for Eswatini (median = 27.3 min, n = 281) and Tanzania (median = 33.3 min, n = 2,678, using full interview datetimes); Nigeria's dataset did not include interview time fields.

Conclusions

The four proposed indicators are practical, transparent, and computable at the record level using standard WHO-VA fields. They enable real-time quality feedback before CoD assignment, do not require large samples, and are compatible with both the 2016 and 2022 WHO-VA instruments. Integration into data management platforms can substantially shorten the feedback loop between data collection and quality improvement, and a reference implementation is publicly available to support adoption.

1. Introduction

Verbal Autopsy (VA) is the primary method for ascertaining causes of death in settings where vital registration systems are absent or incomplete, a situation common to low-resource settings including sub-Saharan Africa, South Asia, and parts of Latin America [1, 2]. A VA interview involves a structured questionnaire administered to a close relative or caregiver of a recently deceased person; the resulting data are used to assign a probable Cause of Death (CoD) through algorithmic methods such as InSilicoVA [3, 4], InterVA [4, 5], SmartVA [4, 6], or physician-coded review [7]. Because these algorithms typically require WHO-VA instrument data reformatted into algorithm-specific input structures, dedicated conversion tools such as pycrossva [8] are commonly used as a preprocessing step, distinct from the data quality assessment that is the focus of this paper.

The quality of VA-derived CoD estimates depends directly on the quality of the underlying VA data. While comprehensive VA data quality assurance encompasses the entire data collection lifecycle (including interview training, instrument design, built-in data validation rule and field supervision), this paper focuses specifically on the assessment of CoD-relevant data quality through quantitative indicators applied to completed interview records. The indicators proposed here are designed to complement, not replace, upstream quality assurance measures by providing systematic, record-level feedback that can identify problems even when training and form-based validation measures have been applied.

Formal quality assessment has typically been embedded in the CoD assignment workflow which includes data analysts or automated algorithms that detect inconsistencies or implausible values when processing aggregated datasets, usually weeks or months after the interviews are conducted [9, 10]. This approach has two key limitations. First, problems identified late are costly to rectify; the interviewer who recorded contradictory responses may no longer be accessible or may have continued the same error across dozens of additional interviews. Second, algorithmic CoD methods require a sufficient number of records to produce stable estimates; a single VA cannot be evaluated for CoD plausibility in isolation. Consequently, individual-record quality problems may go undetected until they affect population-level estimates.

To our knowledge, no published framework provides a standardised set of quality indicators that can be applied to a single VA record, in real time, independently of CoD assignment, and in a format compatible with both major WHO-VA instrument versions. The World Health Organization's Data Quality Review

(DQR) toolkit [11] provides a widely used framework for assessing completeness and internal consistency in routine health information systems, but it operates on aggregated facility- or district-level indicators rather than individual records, and was not designed for the WHO-VA instrument's field structure. Existing quality checks tend to be embedded within specific software packages, are not consistently defined or published, and address only subsets of the data quality problem (for example, completeness alone, or logical range checks only).

We address this gap by proposing four complementary indicators: 1) Informative Completeness Score (ICS), 2) Respondent Reliability Score (RRS), 3) Internal Consistency Index (ICI), and 4) Interview Duration (ID) - summarized at the dataset level as the Median Interview Duration (MID), each targeting a distinct dimension of VA data quality. These indicators can be computed as soon as an interview is complete, providing immediate, record-level feedback to interviewers and supervisors and enabling timely quality evaluation and data-driven decisions before CoD assignment. We demonstrate their application on three national datasets spanning two WHO-VA instrument versions.

2. Methods

2.1 Study Design

This is a methodological paper describing the development and cross-national application of four VA data quality indicators. The indicators are derived from the field structure of the WHO-VA instrument and applied to three existing datasets. No primary data collection was undertaken; accordingly, no ethics approval was required beyond that already governing the source datasets.

2.2 WHO-VA Instrument Compatibility

The WHO publishes the VA instrument in versioned releases; the two most widely used are the 2016 and 2022 versions [12–14]. Both share a common question-naming convention (e.g., id10011 for interview start time, id10019 for deceased sex, id10023 for date of death). The indicators proposed here are constructed from field names present in both versions, making them applicable to data collected with either instrument. Where a field is absent or differently structured in a particular dataset version, computation of the relevant indicator is not possible; this is treated as a data limitation, not an indicator failure. Complete list for indicator field requirement by the WHO VA instrument version is shared in Appendix A1.

2.3 Indicator Definitions

2.3.1 Informative Completeness Score (ICS)

The ICS quantifies the proportion of binary (yes/no) questions (both root questions and their follow-up questions), answered informatively, that is, with a definitive "yes" or "no" rather than "don't know" (dk) or "refused" (ref). Because system-missing fields are excluded from the denominator, the ICS measures the

informativeness of recorded binary responses rather than the completeness of all expected questionnaire fields.

Let B_i denote the set of binary-response fields for record i that received any response (i.e., excluding system-missing values), and let $I_i \subseteq B_i$ be the subset of those fields answered with a definitive "yes" or "no". The ICS for an individual record i is defined as:

$$ICS_i = \frac{|I_i|}{|B_i|} \times 100$$

where $|\cdot|$ denotes set cardinality. ICS ranges from 0% (all responses are "don't know" or "refused") to 100% (all binary responses are informative). A record with ICS = 100% does not imply high accuracy, i.e., a respondent may confidently give wrong answers. But a systematically low ICS signals that the respondent could not or did not engage meaningfully with the interview. ICS is computed over all question fields whose names begin with the standard WHO-VA prefix id, restricting to rows with at least one binary response. Tier classification: ICS: High ($ICS \geq 90$), Moderate (70–89), Low (< 70).

2.3.2 Respondent Reliability Score (RRS)

Respondent reliability and recall accuracy are well-established sources of measurement error in survey data [15], and are particularly salient in VA interviews, where information is reported by a proxy rather than the deceased themselves. In this study, reliability is not measured through repeat interviews, agreement testing or external verification. The RRS should therefore be interpreted as a composite measure of respondent characteristics expected to influence information quality, rather than as a direct measure of response reliability. The RRS for a single record i is a weighted composite of four factors that are empirically associated with the reliability of proxy-reported information: relationship to the deceased, presence at death, recall period, and respondent literacy. For the respondent literacy component, we use the deceased's education variables (id10063 and id10064) as a proxy for respondent education, as the WHO-VA instruments do not capture respondent education directly. The RRS for an individual record i is defined as:

$$RRS_i = W_{rel,i} + W_{prox,i} + W_{rec,i} + W_{edu,i}$$

where each W component is scored as shown in Table 1, and the maximum possible total is 100 points. Each component is evaluated using fields from the same record i , yielding an individual-level score that can be interpreted as a percentage (0-100%).

Table 1
Respondent Reliability Score

Component	Field(s)	Max. Points	Scoring
Relationship (W_{rel})	id10008	40	Spouse/parent/child: 40; family member: 20; other: 10
Presence at death (W_{prox})	id10009	30	Yes: 30; No: 15
Recall period (W_{rec})	id10023, id10012	20	< 90 days: 20; 90–179 days: 15; 180–364 days: 10; ≥365 days: 0
Literacy/education (W_{edu})	id10064, id10063	10	Literate or secondary+ education: 10; No formal education or illiterate: 5
Total		100	Sum of all four components

Because the four components sum to a maximum of 100, RRS can be interpreted as a percentage reliability score, with higher values indicating more reliable respondents. Tier classification is defined as High (RRS ≥ 80), Moderate (50–79), Low (< 50). Records for which the death date or interview date cannot be parsed are excluded from RRS computation, as the recall period component cannot be determined. Like ICS, RRS can be computed for a single record.

2.3.3 Internal Consistency Index (ICI)

The ICI identifies logical contradictions within a VA record where cross-field combinations that are biologically impossible or implausible. For a single record i , define the set of N consistence rules $\mathfrak{R} = \{C_1, C_2, \dots, C_N\}$. Let $V_i \subseteq \mathfrak{R}$ be the subset of rules violated by record i . The ICI for record i is then calculated as:

$$ICI_i = \frac{|\mathfrak{R}| - |V_i|}{|\mathfrak{R}|} \times 100 = \frac{N - |V_i|}{N} \times 100$$

Where $|V_i|$ is the number of violated rules. This produces an individual score from 0% (all N rules violated) to 100% (no violations). A record with any violation receives a penalty proportional to the number of contradictions present. The consistency rules are evaluated per record based on a selected set of variables (Table 2).

It is important to distinguish between inconsistencies that are prevented by built-in data validation rules and skip patterns within the WHO ODK VA questionnaire, and those that are detected by the ICI. Many logical contradictions can be prevented at the point of data entry through proper form design, for example, skip logic can prevent pregnancy-related questions from being asked of male descendants. However, the ICI serves a distinct purpose: it provides an independent assessment of whether these built-in constraints are functioning correctly in practice. Countries may inadvertently modify the standard xForm design or encounter technical issues that break these constraints, and the ICI helps flag such problems. Furthermore, certain inconsistencies such as symptom duration exceeding total illness duration cannot be fully prevented by skip logic and require explicit validation. The ICI therefore assesses both the effectiveness of implemented constraints and the presence of contradictions that slip through

Table 2
Internal Consistency Index (ICI)

Rule	Fields	Condition flagged as violation	Description
C1	id10019, id10305	id10019 = "male" AND id10305 = "yes"	Pregnancy reported for a male decedent
C2	id10153, id10157	id10153 = "no" (no cough) AND id10157 = "yes" (cough up blood)	Blood reported in cough without cough
C3	id10120, id10148	id10148 (fever days) > id10120 (illness days)	Fever duration (days) exceeds total illness duration (days)
C4	id10120, id10154	id10154 (cough days) > id10120 (illness days)	Cough duration exceeds total illness duration
C5	id10120, id10182	id10182 (diarrhoea days) > id10120 (illness days)	Diarrhoea duration exceeds total illness duration
C6	id10120, id10161	id10161 (breathlessness days) > id10120 (illness days)	Breathlessness duration exceeds total illness duration
C7	id10019, id10294, id10304	id10019="male" AND (id10294="yes" OR id10304="yes")	Pregnancy-related symptoms reported for male
C8	id10019, id10344	id10019="male" AND id10344="yes"	Maternal death questions answered for male decedent
C9	id10023, id10012	(id10023) Death date > (id10012) interview date	Interview date precedes death date (temporal consistency)

ICI tier classification is defined as Excellent (ICI $\geq$ 90%), Good (70–89%), Critical (< 70%). At the individual record level, each violation is flagged directly. The ICI is most informative when stratified by interviewer, as systematic violations point to training gaps with a specific interviewer rather than random errors distributed across a team.

2.3.4 Interview Duration (ID)

The Interview Duration (ID) is the elapsed time in minutes between interview start (id10011) and interview end (id10481) for a single record i . ID is computed as:

$$ID_i = \frac{(id10481_i - id10011_i)}{60} \text{ minutes}$$

Records with non-positive or implausible durations (≥ 480 minutes, i.e., 8 hours) are excluded from analysis. The 480-minute threshold was chosen to filter multi-day data entry artefacts while retaining genuinely long interviews.

At the dataset or interviewer level, the primary summary statistic is the Median Interview Duration (MID), the median of all individual ID values. The median is preferred over the mean because ID distributions are typically right-skewed due to outliers and data entry artefacts; the mean and percentile distribution are reported to characterize the full range. This record-level computation enables immediate identification of unusually short interviews that may indicate skipped sections, as well as unusually long interviews that may reflect respondent difficulties or data entry pauses.

ID is computed per record, while MID is computed per dataset. In addition, the MID is computed on all completed VA records in the raw ODK export, including those where consent was not obtained. This is intentional: if an implementation experiences a significant proportion of non-consent cases, these records will be included in the MID computation and will typically exhibit very short durations (as the interview would be terminated early), thereby flagging potential issues with the consent process itself or with interviewer performance in obtaining consent. This allows supervisors to identify systemic consent-related problems that might otherwise go unnoticed.

The fields id10011 and id10481 may store either full ISO 8601 datetimes or time-of-day only (HH:MM:SS), depending on the data collection platform and export configuration rather than on WHO-VA instrument version alone. Where only time-of-day is available, duration computation is unambiguous only when start and end times fall on the same calendar day; an apparent midnight crossing is handled by modular addition of 1,440 minutes, and the 480-minute cap additionally guards against residual cross-day artefacts. Where full datetimes are available, duration is computed directly across date boundaries and this limitation does not apply. In this study, both the Eswatini and Tanzania datasets were extracted with full ISO 8601 datetime precision, so duration was computed directly across date boundaries for both; the time-only fallback path did not apply to either dataset in this analysis.

2.4 Dataset-Level Aggregate Quality Summary Score

For each indicator $Q \in \{ICS, RRS, ICI\}$, the aggregate score for a set of N records is defined as:

$$Q_{dataset} = \frac{1}{N} \sum_{i=1}^N Q_i$$

For MID, the median is the preferred summary statistics due to right-skewed distributions, although the mean can also be reported. These aggregate scores enable supervisors to:

1. Monitor trends over time (e.g., weekly average ICS)
2. Compare interviewers (e.g., average RRS per enumerator)
3. Benchmark sites or regions (e.g., average ICI by districts)
4. Trigger alerts with aggregate score fall below acceptable thresholds

Although each indicator can be calculated for an individual record, the stability of interviewer, site and implementation-level summaries depends on the number and distribution of available records. Small batches may support preliminary monitoring, but comparisons should report sample sizes and uncertainty and should not be treated as stable benchmarks without further evaluation.

2.5 Datasets

Three VA datasets from Tanzania, Eswatini and Nigeria were used to evaluate indicator feasibility and performance (Table 3). For each dataset, we describe the geographic coverage, collection period, sampling or case-identification approach, implementing institutions and level of representativeness. The term “national” is used only where the dataset covers the relevant national VA system rather than a selected implementation, state or surveillance site.

Eswatini (va_2022_es)

A dataset of 380 VA records collected using the 2022 WHO-VA instrument. All four indicators were computable; interview start and end times were stored as full ISO 8601 datetimes with time-zone offsets.

Tanzania (va_2016_tz)

A dataset of 4,344 VA records collected using the 2016 WHO-VA instrument. ICS, RRS, and ICI were computable (483 of 4,344 records were excluded from RRS due to missing or unparseable interview or death dates). Despite the 2016 instrument's nominal time-only specification for interview timing fields, this dataset's id10011 and id10481 values were stored as full ISO 8601 datetimes with time-zone offsets, matching the 2022 datasets; MID was therefore computed directly across date boundaries rather than via the time-only fallback (see Discussion, Section 4.5). The dataset contained duplicate column headers for a subset of fields; the first occurrence was retained.

Nigeria (va_2022_ng)

A dataset of 5,474 VA records collected using the 2022 WHO-VA instrument. ICS and ICI were computable. The fields required for RRS (id10008, relationship to deceased; id10012, interview date) and MID (id10011, interview start time; id10481, interview end time) were absent from this dataset, precluding computation of those two indicators.

2.5 Statistical Analysis

Descriptive statistics (mean, median, standard deviation, minimum, maximum, 5th and 95th percentiles) were computed for ICS and MID (Table 4, Table 5). For RRS, tier distributions (High/Moderate/Low) were reported in addition to continuous summary statistics (Table 6). For ICI, the proportion of records passing all consistency checks and the absolute count of violations per rule were reported. All analyses were performed in Python 3.11 using pandas and NumPy. All indicator computations are designed to be reproducible from the raw CSV exports of WHO-VA instruments. Source code for the python library to calculate the proposed indicators is available on github [https://github.com/vman-tool/vman3_dq]. A compiled version (sandbox) is also available online with sample data for testing purposes [<https://ccva.vatools.net>]

3. Results

3.1 Dataset Overview

Table 3 Dataset characteristics							
Dataset	Country	WHO-VA version	N	Computability			
				ICS	RSS	ICI	MID
va_2022_es	Eswatini	2022	380	Yes	Yes	Yes	Yes
va_2016_tz	Tanzania	2016	4,344	Yes	Yes	Yes	Yes
va_2022_ng	Nigeria	2022	5,474	Yes	No	Yes	No
Note: “Yes” indicates the indicator was computable, and “No” indicates that the required fields were absent from the dataset export							

Table 4 ICS results by dataset							
Dataset	N	Mean (%)	Median (%)	SD	Min (%)	5th pct (%)	95th pct (%)
Eswatini (2022)	380	96.4	98.8	6.6	25.0	86.4	100.0
Tanzania (2016)	4,344	97.4	100.0	7.6	13.4	88.6	100.0
Nigeria (2022)	5,474	99.8	100.0	0.5	88.9	98.8	100.0

Across the three datasets, ICS was uniformly high but varied in dispersion. Tanzania (n = 4,344) showed the greatest variability (SD = 7.6%) and the lowest minimum score (13.4%); Eswatini (n = 380) was intermediate (SD = 6.6%, minimum = 25.0%); and Nigeria (n = 5,474) showed by far the least variability (SD = 0.5%, minimum = 88.9%). Despite this spread in dispersion, central tendency remained consistently high across all three datasets (Tanzania: mean = 97.4%, median = 100.0%; Eswatini: mean = 96.4%,

median = 98.8%; Nigeria: mean = 99.8%, median = 100.0%), indicating that the large majority of respondents engaged informatively with binary questions regardless of implementation approach or instrument version.

The low-end tail (ICS below the 5th percentile, 86.4%) in Eswatini represents 5% of records and may indicate cases where the respondent had limited direct knowledge of the circumstances of death, or where language or comprehension barriers reduced informative response rates. Tanzania's low-end tail is both larger and more informative: its VA program has collected data over a substantially longer period than Eswatini's, and the share of low-ICS records is not evenly distributed across that period. From mid-2024 through late 2025, the monthly proportion of records falling below this threshold repeatedly rose to 6–19%, compared with under 2% in the preceding year, before returning to baseline in the most recent months. This temporal pattern identifies a specific period requiring implementation-level investigation. Possible explanations include changes in supervision, interviewer composition, instrument implementation, respondent characteristics, data migration or export procedures; the present analysis cannot determine the underlying cause.

3.3 Respondent Reliability Score (RRS)

Table 5
RRS continuous summary by dataset

Dataset	N (eligible)	Mean	Median	SD	Min	Max
Eswatini (2022)	349	81.8	87	12.8	35	100
Tanzania (2016)	3,861	84.1	80	13.8	30	100

Table 6
RRS tier distribution by dataset

Dataset	High (≥ 80)	Moderate (50–79)	Low (< 50)
Eswatini (2022)	63.3%	35.0%	1.7%
Tanzania (2016)	67.1%	32.2%	0.6%

RRS was computed for Eswatini (n = 349 of 380; 31 records excluded due to missing or unparseable death date) and Tanzania (n = 3,861 of 4,344; 483 records excluded) (Table 5, Table 6). Across both datasets, the mean RRS exceeded 80 (Tanzania = 84.1, Eswatini = 81.8), placing the average respondent in the High reliability tier.

Tanzania demonstrated a higher proportion of High-tier respondents (67.1% vs 63.3%) and a higher mean RRS. Its median, however, was lower than Eswatini's, indicating a longer lower tail in the Tanzania distribution despite its higher average. The Low-tier proportion remained small in both datasets (Tanzania = 0.6%, Eswatini = 1.7%), suggesting that respondents with very low reliability scores are uncommon in well-managed programs.

The Nigeria dataset was excluded from RRS computation because the relationship-to-deceased field (id10008) and interview date field (id10012) were absent from the exported data. This illustrates that RRS requires specific fields that may not be available in all dataset exports and highlights the importance of ensuring data export completeness.

3.4 Internal Consistency Index (ICI)

The ICI and per-rule violation counts are presented in Table 7.

Table 7
ICI results and rule-level violation counts

Rule	Description	Tanzania (2016) (n = 4344)	Eswatini (2022) (n = 380)	Nigeria (2022) (n = 5474)
C1	Pregnancy reported for a male decedent	0 (0.0%)	0 (0.0%)	1 (0.0%)
C2	Blood reported in cough without cough	0 (0.0%)	0 (0.0%)	0 (0.0%)
C3	Fever duration exceeds total illness duration	42 (1.0%)	4 (1.1%)	115 (2.1%)
C4	Cough duration exceeds total illness duration	71 (1.6%)	23 (6.1%)	10 (0.2%)
C5	Diarrhoea duration exceeds total illness duration	15 (0.3%)	7 (1.8%)	n/a (field absent)
C6	Breathlessness duration exceeds total illness duration	27 (0.6%)	8 (2.1%)	12 (0.2%)
C7	Pregnancy-related symptoms reported for a male decedent	30 (0.7%)	0 (0.0%)	323 (5.9%)
C8	Maternal death questions answered for a male decedent	0 (0.0%)	0 (0.0%)	0 (0.0%)
C9	Interview date precedes death date (temporal impossibility)	152 (3.5%)	0 (0.0%)	n/a (field absent)
Records passing all applied rules		92.8%	91.8%	92.0%
Mean errors per record		0.078	0.111	0.084

Eswatini achieved an ICI of 91.8% (Excellent tier), Tanzania 92.8% (Excellent), and Nigeria 92.0% (Excellent) across the nine consistency rules (Table 7). No dataset fell below the Good threshold (70%). The most striking single-rule finding remains the pattern of C4 violations in Eswatini: 23 records (6.1%) had a cough duration exceeding the reported total illness duration, compared with 71 records in Tanzania's much larger dataset (1.6%) and only 10 in Nigeria (0.2%). This discrepancy may reflect a data entry convention in the Eswatini dataset, a structural difference in how interviewers were trained to elicit

duration responses, or ambiguity in how respondents distinguished disease-specific symptom duration from total illness episode duration.

Pregnancy-related symptoms reported for a male decedent (C7) was flagged in 323 Nigerian records (5.9%) by far the largest single-rule violation count of any dataset compared with 30 records in Tanzania (0.7%) and none in Eswatini. Field-level review confirmed that the underlying questions correspond to pregnancy-related questions in the Nigerian instrument. However, further examination of the form configuration and complete data pipeline is required to determine whether the observed values arose during interview administration, form customization, default coding, data transformation or export. The pattern therefore warrants investigation into whether the male/female skip logic is functioning correctly in that implementation's data pipeline. Separately, C9 (interview date preceding death date) was violated in 152 Tanzanian records (3.5%). This is a rate not observed at all in Eswatini and not computable in Nigeria, where the interview-date field is absent from the export. This pattern is temporally implausible by definition and, taken together with the concentration of low-ICS records in Tanzania during mid-2024 through late 2025, is consistent with a period of reduced data-entry oversight during program implementation rather than an isolated measurement issue.

The near-absence of C1 (pregnancy in male decedent) and C8 (maternal-death questions for male decedent) violations across all datasets suggests that the skip logic in the WHO-VA instrument generally functions correctly in preventing these questions from being asked of inappropriate respondents. However, the elevated C7 violations in Nigeria demonstrate that skip logic can be compromised (possibly through form customization, data export issues, or interviewer behavior) and that the ICI provides an important independent check. This reinforces the value of the ICI as a tool for detecting when built-in constraints have been inadvertently broken, which might otherwise remain invisible until CoD assignment produces implausible results.

The C1 rule (pregnancy in a male decedent) was virtually absent across all datasets (only 1 case, in Nigeria), consistent with this being a clear biological impossibility that interviewers and data entry systems readily avoid. C2 (blood reported without cough) and C8 (maternal-death questions answered for a male decedent) were similarly absent across all three datasets, suggesting the underlying skip logic for these implausible combination functions reliably across both instrument versions and all three data pipelines.

The C3 violation (fever duration > illness duration) remained most prevalent in Nigeria (n = 115, 2.1%), pointing to a potential systematic data entry pattern worth investigating at the interviewer level. When stratified by the interviewer, the ICI can be used to reveal whether these violations (and the C7 violations noted above) are concentrated in specific enumerators or sites.

3.5 Median Interview Duration (MID)

Table 8
MID results by dataset

Dataset	N (valid)	Mean (min)	Median (min)	SD	5th pct (min)	95th pct (min)
Eswatini (2022)	281	56.9	27.3	79.1	11.7	235.1
Tanzania (2016)	2,678	63.1	33.3	81.2	12.8	247.9

MID results are presented in Table 8. Both Eswatini's and Tanzania's MID were computed from full ISO 8601 datetime fields. The median interview duration in Eswatini (27.3 minutes, n = 281) is consistent with expected VA interview lengths for a standard instrument administered by trained interviewers. The high mean (56.9 minutes) and wide standard deviation (79.1 minutes) reflect a right-skewed distribution with a substantial proportion of records reporting very long durations (95th percentile = 235.1 minutes). In datasets storing full datetimes, multi-day intervals (e.g., where the interview was paused and resumed the following day) can produce apparent durations of many hours; these are captured before the 480-minute exclusion and inflate the mean.

Tanzania's MID distribution (n = 2,678) follows the same right-skewed pattern as Eswatini's, with every summary statistic shifted higher: median = 33.3 minutes (vs Eswatini's 27.3), mean = 63.1 minutes (vs 56.9), SD = 81.2 minutes (vs 79.1), and 95th percentile = 247.9 minutes (vs 235.1). As in Eswatini, this reflects genuine multi-hour or multi-day intervals (e.g., paused-and-resumed interviews) that inflate the upper tail before the 480-minute exclusion is applied. That both datasets show a mean roughly double their median reinforces the manuscript's guidance to report median rather than mean as the preferred MID summary statistic.

Nigeria's dataset did not include the interview start and end time fields (id10011, id10481), preventing MID computation. Programs using the 2022 WHO-VA instrument that wish to monitor MID should verify that these fields are included in data exports. In these two datasets, approximately 12 minutes represented the lower fifth percentile of interview duration. This exploratory, dataset-specific threshold may be used to identify records for review, but it should not be treated as a general quality standard until its association with missingness, inconsistencies, or independently assessed interview quality has been established. Similarly, records with MID above the 95th percentile may represent multi-session interviews or data entry artefacts.

4. Discussion

4.1 Principal Findings

We have proposed and applied four complementary VA data quality indicators that can be computed from a single interview record, without requiring CoD assignment or aggregated data. The indicators were applicable to data collected with both the 2016 and 2022 WHO-VA instruments and performed

consistently across datasets from three countries. Across all datasets where indicators were computable, the majority of records achieved satisfactory performance (ICS > 94%, RRS High tier, ICI pass), confirming that these datasets represent well-conducted field implementations. At the same time, each indicator identified a non-trivial subset of records or patterns worthy of follow-up, demonstrating that the indicators add value even in high-quality programs.

4.2 Advantages Over Post-Hoc Quality Assessment

The central argument for this indicator framework is temporal: where quality problems identified before CoD assignment can still be corrected. A supervisor alerted in real time that a specific interviewer is generating a disproportionate number of C4 violations (cough duration > illness duration) can provide targeted retraining. A data manager notified that 15% of today's uploads have ICS below 80% can investigate the cause (i.e., interviewer fatigue, a difficult community, or a software display problem) before the situation worsens. In contrast, quality issues identified during CoD coding may require the original respondent to be re-contacted, which is often logistically impossible months after the interview. Some errors, such as contradictory duration fields, may be structurally unresolvable and require the record to be flagged or excluded. Moreover, unlike CoD assignment algorithms which require a population distribution to produce meaningful estimates (e.g., InSilicoVA requires at least 100 cases for stable output [3]), the indicators proposed here are fully interpretable at the level of a single record. A VA data collection implementation with 10 records can assess ICS, RRS, and ICI immediately; it need not wait for a sufficient sample to run a CoD algorithm.

It is important to clarify that this framework addresses the assessment of CoD-relevant data quality, complementing (rather than replacing) upstream quality assurance measures such as interviewer training, instrument design, and built-in validation rules. While these upstream measures are essential for preventing errors, the indicators proposed here provide a systematic means of detecting errors that persist despite such measures, or that arise from unintended modifications to the data collection process.

4.3 Complementarity of the Four Indicators

The four indicators measure distinct aspects of quality that are not reducible to one another. A record can score perfectly on ICS (all binary questions answered definitively) yet still contain internal contradictions (low ICI) because the respondent confidently gave inconsistent answers. An interview can be internally consistent yet reflect an unreliable respondent (low RRS). For example, a distant acquaintance was interviewed years after the death. An interview can be conducted by a highly reliable respondent and contain no logical errors, yet be completed in three minutes, suggesting that key sections were skipped (short ID, lowering the overall MID). The simultaneous availability of all four indicators gives supervisors a multidimensional view of record quality.

4.4 Dataset-Specific Observations

The absent fields in Nigeria's dataset (id10008, id10011, id10012, id10481, and id10182) underscore that the indicator framework's utility depends on data export completeness. Field absence is itself a data quality signal: if an implementation is not capturing interview timestamps or diarrhoea duration, the MID indicator and one of ICI's duration-based rules cannot serve their monitoring function. Programmers should ensure that all instrument fields, including administrative, timing, and symptom-duration fields, are exported alongside the interview responses.

The anomalously high C4 violation rate in Eswatini compared to Tanzania and Nigeria (6.1% vs 1.6% and 0.2%, respectively) illustrates that ICI performance is context-dependent. Possible explanations include: (1) interviewer training that did not sufficiently emphasize the distinction between symptom onset and total illness duration; (2) cultural or linguistic differences in how respondents conceptualize "illness duration" versus "symptom duration"; or (3) data entry conventions specific to the Eswatini VA implementation. Investigating this pattern at the interviewer level (by stratifying ICI by interviewer ID) would be the appropriate next step.

In addition, the Nigeria dataset showed a markedly elevated C7 violation rate (pregnancy-related symptoms reported for a male decedent; 5.9%, versus 0.7% in Tanzania and none in Eswatini). The underlying fields correspond to pregnancy-related questions in the standard WHO-VA instrument, indicating this reflects a program-specific weakness in the male/female skip logic applied during data collection or export, rather than a field-mapping artifact during analysis. This finding is particularly instructive because it demonstrates how the ICI can detect failures in built-in form constraints either through form customization that inadvertently disabled skip logic, or through data export issues that led to pregnancy-related responses being recorded for male descendants. Such failures might otherwise remain invisible until CoD assignment produces implausible cause-of-death patterns.

Tanzania showed an elevated C9 violation rate (interview date, id10012, preceding death date, id10023; 3.5%, versus 0.0% in Eswatini and not computable in Nigeria, where id10012 is absent). This is consistent with the concentration of low-ICS records observed in Tanzania during mid-2024 through late 2025 discussed above, together pointing to a discrete period of reduced data-entry oversight in the VA data collection implementation rather than a persistent, instrument-wide problem.

4.5 Instrument Compatibility

A strength of this framework is compatibility with both the 2016 and 2022 WHO-VA instruments. The field names used for each indicator are defined consistently across versions. The MID computation accommodates both full ISO 8601 datetime values and time-only (HH:MM:SS) values for the interview start/end fields, applying a modular adjustment for an apparent midnight crossing when only time-of-day is available. This adaptability matters in practice: whether these fields carry full datetime or time-only precision depends on how the ODK form configures the corresponding question (a date/Time widget versus a time widget) and on whether any downstream export or processing step truncates that precision before analysis, a pattern consistent with broader documented data-quality variability across mobile data collection platforms used in low- and middle-income country health surveillance [16, 17].

The Tanzania dataset validated in this analysis was retrieved directly from the ODK Central server and retained full datetime precision throughout; an earlier analysis, using roughly six years of accumulated historical Tanzania data from an ODK aggregate server, showed only time-of-day precision for the same fields, an artefact of an ODK form-variable configuration, an intermediate export or processing step rather than a property of the 2016 instrument itself. Programmers should therefore both configure the relevant ODK fields to capture full datetime and confirm that no step in their data pipeline degrades that precision before it reaches analysis. Where only time-of-day is available, cross-midnight interviews cannot be distinguished from negative durations and are excluded, and median rather than mean duration should be reported. Researchers considering application of these indicators to earlier instrument versions (pre-2016) should verify field name correspondence, as the question numbering convention evolved substantially before the 2016 release.

4.6 Limitations

Several limitations should be acknowledged. First, this study evaluates the feasibility, distribution and cross-dataset applicability of the proposed indicators but does not establish criterion validity against an independent measure of VA record quality. The indicators were not compared with re-interview findings, expert record review, audio audits, cause-of-death concordance or another external reference standard. Accordingly, the findings should be interpreted as an initial methodological evaluation rather than definitive validation. Second, the thresholds defining ICS, RRS, and ICI tiers (e.g., ICI Excellent $\geq 90\%$, RRS High ≥ 80) are expert-derived and have not yet been empirically validated against CoD concordance data. Future work should assess whether records in different ICS or RRS tiers show systematically different CoD assignment outcomes in comparison with gold-standard CoD data.

Third, the ICI rule set (C1–C9), while now covering implausible pregnancy/maternal responses for male decedents (C1, C7, C8) alongside symptom-duration and temporal-consistency checks, still represents a subset of possible logical inconsistencies in the WHO-VA instrument; it does not yet include internal-consistency checks specific to genuine female maternal deaths, nor coverage of the pediatric and neonatal sections or age-sex combinations for other conditions. Relatedly, the exact field semantics behind newer rules such as C7 and C8 were confirmed against the instrument for the datasets in this study, but researchers adapting this framework to new VA implementation programs should not assume field-ID groupings alone are sufficient and should verify question-level semantics directly.

Fourth, RRS is limited to datasets that capture respondent characteristics. Programs that do not record respondent relationships or death date will find this indicator inapplicable. In addition, the RRS's education/literacy component relies on a proxy measure: neither the 2016 nor the 2022 WHO-VA instrument captures respondent education directly; both capture the deceased's education level (fields id10063 and id10064). We therefore use the deceased's education as a proxy for respondent education, a choice supported by evidence of educational homogamy where in many settings, the respondent (often a spouse or adult child) is likely to have educational attainment comparable to that of the deceased [18, 19]. However, where the respondent is not a close family member, this proxy may introduce measurement error. We recommend that future revisions of the WHO-VA instrument consider

adding a dedicated field for respondent education level to enable more accurate calculation of the RRS and to support broader analyses of respondent-related bias in VA data.

Fifth, the datasets used for validation are not random samples and may not represent the full range of data quality conditions in their respective countries.

Sixth, indicator values (particularly MID) can be sensitive to the specific data export or processing pipeline used to generate them, independent of the underlying instrument version or VA implementation data quality: two extractions of the same country's data six years apart, one retrieved directly from the ODK Central server and one from ODK Aggregate with intermediate export pipeline, showed substantially different timestamp precision for the same fields. Indicator results should therefore be interpreted as properties of a specific, documented data pull rather than fixed properties of the VA implementation.

Seventh, as noted in the methods, these indicators are computed on raw ODK export data, including records where consent was not obtained. This is by design, as it allows the indicators to flag problems with the consent process itself, for example, a high proportion of very short duration interviews may indicate that interviewers are struggling to obtain consent and terminating interviews prematurely. However, it also means that programs using these indicators should be aware that records with missing or incomplete data due to non-consent will affect aggregate scores and should be interpreted accordingly.

4.7 Future Directions

The indicator framework is intended to be extensible, and its reference implementation (the `vman_dq` Python package used throughout this study) is publicly available on GitHub, facilitating both independent validation and integration beyond the `vman3` platform. Future work should: (1) validate tier thresholds against linked CoD data; (2) expand the ICI rule set to cover the pediatric, neonatal, and under-5 sections of the WHO-VA instrument, and add internal-consistency checks specific to genuine female maternal deaths, building on the male-decedent pregnancy/maternal-module checks (C7, C8) introduced in this analysis; (3) investigate machine-learning extensions that learn anomalous response patterns from historical data and flag records that deviate from expected distributions; (4) develop composite data quality scores that aggregate across the four indicators; (5) integrate the indicators into field data collection platforms to provide immediate, enumerator-facing feedback, an effort the public availability of the reference implementation is intended to support; and (6) establish platform-specific guidance for maintaining timestamp precision across data export pipelines. This study's Tanzania data illustrates the risk concretely: an extraction from ODK Aggregate (the legacy ODK server software) over six years prior stored interview start/end times as time-of-day only, whereas the current extraction, from ODK Central (its successor), retained full datetime precision for the same fields. As many national VA programs hold historical archives predating their migration from Aggregate to Central, analyses spanning both eras should explicitly account for this source of heterogeneity in timestamp precision.

5. Conclusion

We propose four complementary indicators (ICS, RRS, ICI, and MID) that collectively address completeness, respondent reliability, logical consistency, and interview process quality in VA data. Unlike existing approaches that assess quality as a byproduct of CoD assignment, these indicators are designed for real-time application to individual records, enabling supervisors to identify and act on quality problems during data collection rather than after it. The indicators are compatible with both the 2016 and 2022 WHO-VA instruments, require no additional data beyond the standard instrument, and can be computed automatically by data management systems.

Cross-national application to datasets from Eswatini, Tanzania, and Nigeria demonstrated the feasibility of calculating the proposed indicators across two WHO-VA instrument versions and identified program-specific patterns requiring follow-up. Direct comparison across countries should nevertheless be cautious because not every indicator or consistency rule was evaluable in every dataset, sample sizes differed substantially, and the datasets were not selected to be nationally representative. The elevated cough-duration inconsistency in Eswatini, pregnancy-related response inconsistencies in Nigeria, and a discrete period of reduced data-entry oversight in Tanzania visible only once its full interview-timing data became available. That these patterns emerged directly from routine indicator computation, rather than from a dedicated investigation, illustrates the framework's practical value: it surfaces actionable, interviewer- or program-level quality issues as a routine byproduct of data collection rather than requiring a separate audit.

We emphasize that this framework complements rather than replaces the upstream quality assurance measures. By providing systematic, quantitative feedback on completed records, these indicators enable VA programs to detect problems that persist despite training and form-based validation, including issues arising from unintended modifications to standard instruments. The indicators also provide a means of monitoring the effectiveness of built-in data constraints and identifying when they have been inadvertently broken.

We encourage adoption of these indicators as a standard component of VA data management workflows. To support this, the reference implementation used throughout this study is made publicly available on GitHub as an open-source Python package, allowing VA programs to adopt, validate, and extend the framework independently of any single data management platform.

Declarations

Conflict of interest

The author declares no conflict of interest.

Funding

No specific funding was received for this methodological study.

Author Contribution

IL conceptualized the indicator framework, developed the computational algorithms, performed the validation analysis, and drafted the initial manuscript. SI developed the online sandbox environment for testing and analysis, and contributed to the manuscript revision. GM (Ministry of Health, Tanzania), MM (Ministry of Health, Eswatini), BZ (Clinton Health Access Initiative, Nigeria), and SY (Katsina State Primary Health Care Agency, Nigeria) reviewed the manuscript, confirmed the results, and ensured alignment with their respective country datasets and programmatic contexts. MT and RW provided overall oversight, critically reviewed the manuscript for intellectual content, and contributed to the final validation of the findings. All authors read and approved the final manuscript.

Data Availability

The indicator computation algorithms used in this study are openly available. The Python package `vman3-dq` is available on GitHub at https://github.com/vman-tool/vman3_dq and as a compiled demonstration sandbox at <https://ccva.vatools.net>. These resources allow for independent validation, adaptation, and application of the proposed framework. The individual-level verbal autopsy datasets analysed during the current study are not publicly available due to restrictions imposed by the respective data custodians to protect participant privacy and confidentiality. These datasets were obtained from the national Ministries of Health of Tanzania, Eswatini, and Nigeria under data-sharing agreements that prohibit redistribution. Access to the data may be requested from the corresponding ministries:• Tanzania: Ministry of Health, Tanzania ([relevant contact details or statement that requests are handled through institutional processes])• Eswatini: Ministry of Health, Kingdom of Eswatini• Nigeria: Clinton Health Access Initiative (CHAI), Nigeria, and Katsina State Primary Health Care AgencyRequests for access to these datasets should be submitted to the relevant ministry or data custodian. De-identified data may be available to qualified researchers upon reasonable request and with the explicit permission of the respective data custodians.

References

1. Mikkelsen L, Phillips DE, AbouZahr C, et al. A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *Lancet*. 2015;386(10001):1395–406.
2. Fottrell E, Byass P. Verbal autopsy: methods in transition. *Epidemiol Rev*. 2010;32:38–55.
3. McCormick TH, Li ZR, Calvert C, et al. Probabilistic Cause-of-death Assignment using Verbal Autopsies. *J Am Stat Assoc*. 2016;111(515):1036–49.
4. Li ZR, Thomas J, Choi E, et al. The openVA Toolkit for Verbal Autopsies. *R J*. 2022;14(4):316–34.
5. Byass P, Calvert C, Miiro-Nakiyingi J, et al. InterVA-4 as a public health tool for measuring HIV/AIDS mortality: a validation study from five African countries. *Glob Health Action*. 2013;6:22448.

6. Serina P, Riley I, Stewart A, et al. A shortened verbal autopsy instrument for use in routine mortality surveillance systems. *BMC Med.* 2015;13:302.
7. Lozano R, Lopez AD, Atkinson C, et al. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr.* 2011;9:32.
8. verbal-autopsy-software/pyCrossVA [Internet], Verbal S. 2026 [cited 2026 Aug 2]. Available from: <https://github.com/verbal-autopsy-software/pyCrossVA>
9. Sankoh O, Byass P. The INDEPTH Network: filling vital gaps in global epidemiology. *Int J Epidemiol.* 2012;41(3):579–88.
10. Oza S, Lawn JE, Hogan DR, et al. Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013. *Bull World Health Organ.* 2015;93(1):19–28.
11. Data quality assurance (DQA.) [Internet]. [cited 2026 Aug 2]. Available from: <https://www.who.int/data/data-collection-tools/health-service-data/data-quality-assurance-dqa>
12. World Health Organization. Verbal autopsy standards: The 2016 WHO verbal autopsy instrument [Internet]. Geneva: WHO. 2016 [cited 2024 Dec 9]. Available from: <https://www.who.int/publications/m/item/verbal-autopsy-standards-the-2016-who-verbal-autopsy-instrument>
13. World Health Organization. Verbal autopsy standards: The 2022 WHO verbal autopsy instrument [Internet]. Geneva: WHO. 2022 [cited 2026 July 31]. Available from: <https://www.who.int/publications/m/item/2022-who-verbal-autopsy-instrument>
14. Nichols EK, Byass P, Chandramohan D, et al. The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLOS Med.* 2018;15(1):e1002486.
15. Saris WE. Total Survey Error. *Encycl Qual Life Well- Res* [Internet]. Springer, Dordrecht; 2014 [cited 2026 Aug 2]. pp. 6703–6704. Available from: https://link.springer.com/rwe/10.1007/978-94-007-0753-5_3034
16. Marks M, Lal S, Brindle H et al. Electronic Data Management for Vaccine Trials in Low Resource Settings: Upgrades, Scalability, and Impact of ODK. *Front Public Health* [Internet]. 2021 [cited 2026 Aug 2];9.
17. Olu-Abiodun O, Fatureti A, Adepoju A, et al. Effectiveness and challenges of digital tools implementation for enhancing infectious disease surveillance data quality in low- and middle-income countries: A systematic review protocol. *PLoS ONE.* 2025;20(8):e0330904.
18. Schwartz CR. Trends and Variation in Assortative Mating: Causes and Consequences. *Annu Rev Sociol.* 2013;39:451–70.
19. Eika L, Mogstad M, Zafar B. Educational Assortative Mating and Household Income Inequality. *J Polit Econ.* 2019;127(6):2795–835.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendix.docx](#)