NIJAMSS Page: 49-55

NEXUS INTERNATIONAL JOURNAL OF APPLIED SCIENCE, MEDICINE AND SOCIAL SCIENCES

Applied Methodological Review
Volume: 01
Issue: 01

Open Access | www.nijamss.org | Vol. 1, Issue 1, May 2026 | Paper No. 008

FROM RAW OBSERVATIONS TO RELIABLE EVIDENCE: A PRACTICAL FRAMEWORK FOR DATA CLASSIFICATION, TABULATION AND STATISTICAL INTERPRETATION

Dr. Daniel A. Otwori, PhD

School of Public Health, University of East Africa Baraton, Kenya

Thematic Area: Biostatistics, Data Analytics and Research Methods | Manuscript Type: Applied Methodological Review

Article Received: 4th May 2026, Published on: 31st May 2026

***Corresponding Author: Dr. Daniel A. Otwori**
School of Public Health, University of East Africa Baraton, Kenya
Email: otworid@uaeb.ac.ke

ABSTRACT

Data classification and tabulation are foundational stages in statistical analysis because they determine how raw observations are represented, summarized, compared and interpreted. Yet these stages are frequently treated as mechanical preliminaries rather than as analytical decisions with direct consequences for validity. This applied methodological review examines the classification of qualitative and quantitative data, the design of categories, the construction of frequency and contingency tables, and the risks introduced by inappropriate aggregation. Drawing on biostatistical, epidemiological and data-management literature, the paper shows that accurate classification requires alignment among the research question, measurement scale, variable coding and planned statistical analysis. Nominal, ordinal, interval and ratio variables support different summaries and inferential procedures; discrete and continuous variables also require different handling. Tabulation improves transparency by revealing distributions, missingness, subgroup differences and possible data errors, but poorly designed tables can conceal sparse cells, merge clinically distinct groups, exaggerate patterns through inappropriate percentages or suppress meaningful variation. The review proposes a Classification-Tabulation-Interpretation Quality Framework built around seven principles: conceptual validity, mutually exclusive and exhaustive categories, preservation of information, denominator transparency, missing-data visibility, reproducibility and ethical disaggregation. Three applied tables demonstrate how to select data structures, construct interpretable tabulations and audit table quality. The paper concludes that classification and tabulation are not merely presentation tasks; they are core components of statistical reasoning. Strengthening training, documentation, metadata standards and software-assisted quality checks can improve the reliability, comparability and policy relevance of research evidence.

KEYWORDS: Data classification; tabulation; biostatistics; frequency distribution; contingency table; data quality; research interpretation.

I. INTRODUCTION

Statistical analysis begins before a model is fitted or a hypothesis is tested. Its quality depends fundamentally on how observations are defined, coded, grouped, and presented. Classification transforms raw observations into analytically meaningful variables, whereas tabulation arranges those variables systematically so that patterns, differences, errors, and uncertainty become visible (Bernardi et al., 2023).

In public health, inappropriate classification can have significant practical consequences. Combining distinct disease categories may conceal variations in risk, while excessively broad age or income groupings can obscure vulnerable populations. Likewise, tables presenting percentages without clearly defined denominators may misrepresent disease burden or programme performance. Statistical software cannot independently correct these weaknesses because it applies the definitions, categories, and decision rules provided by the analyst. Consequently, misclassification and incomplete health records can bias statistical estimates and conclusions (Penev et al., 2024).

Contemporary datasets are increasingly large, linked, heterogeneous, and multidimensional. Routine health information systems, surveys, registries, electronic medical records, and administrative databases frequently integrate variables collected for different purposes (Janssen et al., 2024). This complexity increases the need for explicit coding rules, standardised metadata, reproducible workflows, and transparent tabulation (Finster et al., 2025). Accordingly, this paper treats classification and tabulation as integral components of data quality, statistical validity, and responsible evidence use.

A. Problem Statement

Research reports commonly describe data types and table formats, yet insufficient attention is given to the analytical consequences of classification and tabulation decisions. Categories may overlap, exclude valid observations, conceal clinically important variations, or change across data-collection periods. Similarly, arbitrary grouping of continuous variables may cause information loss, reduce statistical precision, and produce unstable findings. Missing values may also be omitted from tables without explanation, making percentages appear more complete and reliable than the underlying data warrant (Bernardi et al., 2023; Penev et al., 2024).

The central methodological problem is that classification decisions are frequently made after data collection, without adequate alignment with theory, measurement quality, or the intended statistical analysis. Tabulation subsequently reproduces these weaknesses in a visually authoritative form. Because readers often regard tables as objective summaries of evidence, errors in their construction may distort interpretation, influence policy decisions, and compromise subsequent analyses, even when the numerical calculations are technically correct. Therefore, transparent coding procedures, explicit denominators, appropriate category boundaries, and complete reporting of missing data are essential for ensuring valid and reproducible findings (Finster et al., 2025).

B. Aim, Objectives and Research Questions

The aim of this paper is to critically examine the principles and practical methods of data classification and tabulation and to develop an applied quality framework for statistical research.

i. Clarify the major classifications and measurement scales used in biostatistics and applied research.
ii. Explain the principles of valid category design, coding and metadata documentation.
iii. Describe the process of constructing frequency, cross-tabulation and contingency tables.
iv. Evaluate how classification and tabulation affect statistical interpretation, comparability and inferential validity.
v. Propose practical quality checks and recommendations for researchers, students and institutions.

The principal question is: How can data classification and tabulation be designed to preserve information, improve transparency and support valid statistical interpretation?

II. MATERIALS, METHODS AND TECHNICAL PROCEDURES

A. Review Design

An applied methodological review was conducted to synthesise established statistical principles and contemporary data-management practices. It examined variable types, measurement scales, category construction, table design, missing-data reporting, reproducibility, and accurate interpretation of findings (Bernardi et al., 2023; Penev et al., 2024; Finster et al., 2025)..

B. Evidence Sources

Sources included authoritative biostatistics and categorical-data texts, research-design references, official data-quality and reporting guidance, and literature on reproducible and FAIR data practices. Foundational sources were retained because the relevant mathematical principles remain stable, while recent data-governance evidence strengthened the discussion of metadata, interoperability, standardisation, and responsible disaggregation (Wilkinson et al., 2016; Penev et al., 2024; Finster et al., 2025).

C. Analytical Procedure

Evidence was organised into four stages: defining the construct; assigning an appropriate measurement scale and coding system; tabulating data with transparent denominators; and interpreting patterns while considering missingness, uncertainty, and study design. Practical examples demonstrated how identical datasets may produce different conclusions depending on the classification and tabulation choices applied (Bernardi et al., 2023).

D. Quality Criteria

Classification and tables were evaluated against seven criteria: conceptual validity, mutual exclusivity, exhaustiveness, preservation of useful information, transparent denominators, visible missing data, and reproducibility. Ethical disaggregation was incorporated as a cross-cutting criterion to ensure that subgroup reporting advances equity without increasing confidentiality or disclosure risks (Penev et al., 2024).

E. Limitations

This paper is a methodological synthesis and does not estimate the prevalence of classification or tabulation errors in published research. Examples are illustrative rather than derived from a single empirical dataset. Software-specific

procedures may also vary, although the underlying principles apply across Excel, SPSS, Stata, R, Python and other analytical platforms.

III. RESULTS AND TECHNICAL FINDINGS

A. Data Types and Measurement Scales

Data are broadly classified as categorical or numerical, but valid analysis requires precise identification of their measurement scales (Mishra et al., 2019). Nominal variables comprise unordered categories; ordinal variables have a meaningful order but unequal or unknown intervals; interval variables have equal units without a true zero; and ratio variables have equal units and a meaningful zero. Discrete variables assume countable values, whereas continuous variables can take any value within a defined range (Penev et al., 2024).

Table 1. Data classifications, examples and appropriate summaries

Data class	Measurement scale	Example	Permissible ordering	Preferred summaries	Common analytical methods
Categorical	Nominal	Sex recorded for analysis, blood group, county	No inherent order	Counts, proportions and mode	Chi-square, Fisher's exact test, logistic models
Categorical	Ordinal	Disease stage, satisfaction level, wealth quintile	Ordered categories	Counts, proportions, median category and cumulative percentages	Rank tests, ordinal regression
Numerical	Discrete ratio	Number of clinic visits or children	Ordered, countable values	Mean, median, range and variance	Poisson/negative binomial models
Numerical	Continuous interval	Temperature in degrees Celsius	Ordered with equal units	Mean, standard deviation, percentiles	Correlation, linear models and time-series methods
Numerical	Continuous ratio	Age, weight, income and blood pressure	Ordered with equal units and true zero	Mean/SD or median/IQR depending on distribution	t tests, ANOVA, regression and survival analysis
Time-to-event	Duration with censoring	Time to recovery, death or relapse	Ordered time with incomplete observation possible	Median survival and survival probabilities	Kaplan-Meier and Cox regression

B. Principles of Valid Classification

Classification should begin with the research construct rather than with available software options. A valid A sound classification represents the concept being studied, uses non-overlapping categories, and provides a coding pathway for every legitimate observation. Categories should remain stable across observers and time, while any changes must be documented clearly (Bernardi et al., 2023; Penev et al., 2024).

i. **Conceptual validity:** Categories must correspond to the study construct and analytical purpose.

ii. **Mutual exclusivity:** An observation should not be assigned to multiple categories unless multiple-response coding is intended.

iii. **Exhaustiveness:** Every valid response should have a coding pathway, including clearly defined "other" or "not applicable" options where necessary.

iv. **Consistency:** Definitions and coding rules should be applied uniformly across sites, personnel, and data-collection periods.

v. **Parsimony:** Categories should preserve meaningful variation without becoming so numerous that resulting tables are sparse or unstable.

vi. **Reversibility:** Original values should be retained whenever possible, enabling alternative groupings and sensitivity analyses (Wilkinson et al., 2016).

vii. **Metadata documentation:** Labels, units, coding rules, derived variables, and category changes should be recorded systematically in a comprehensive data dictionary (Finster et al., 2025).

C. Risks of Categorising Continuous Variables

Categorising a continuous variable may simplify communication but inevitably reduces information. Reporting age in broad groups, for example, conceals within-group variation and creates artificial differences around category boundaries. Data-driven cut-off points may also inflate statistical significance, reduce statistical power, and weaken reproducibility (Altman & Royston, 2006; Bennette & Vickers, 2012).

Where categorisation is necessary for clinical or policy interpretation, cut-off points should follow established standards, substantive knowledge, or prespecified rules. Analyses should preferably retain continuous measurements, using grouped results mainly for descriptive reporting or sensitivity analysis (Royston et al., 2006).

D. The Tabulation Process

Tabulation should follow a documented sequence: verify coding; define the table population; select rows and columns; determine whether to present counts, row percentages, column percentages, or rates; display missing values; calculate totals; check internal consistency; and provide an informative title and explanatory footnotes (Bernardi et al., 2023; Penev et al., 2024). Each table should address a specific analytical question rather than display every available variable.

Table 2. Main table types and their analytical purposes

Table type	Primary purpose	Typical contents	Key denominator	Main interpretation risk
Simple frequency table	Describe one categorical variable	Counts and percentages by category	All valid observations or total sample	Missing values omitted or percentages based on unclear total
Grouped frequency distribution	Summarise a numerical distribution	Intervals, frequencies, cumulative frequencies and percentages	All observations in the stated range	Unequal interval widths or arbitrary cut-points
Two-way contingency table	Compare two categorical variables	Cell counts with row or column percentages	Row total, column total or full table total	Using the wrong percentage direction
Multi-way cross-tabulation	Examine stratified patterns	Two variables displayed within a third grouping variable	Stratum-specific totals	Sparse cells and over-interpretation
Summary-statistics table	Compare numerical outcomes across groups	Mean/SD, median/IQR, range and sample size	Non-missing observations for each variable	Applying mean/SD to highly skewed data without explanation
Rate or risk table	Compare occurrence across populations or time	Events, population/person-time and rate or risk	Population at risk or person-time	Numerator-denominator mismatch
Missing-data table	Assess completeness and possible bias	Missing counts and percentages by variable or subgroup	Expected observations	Treating missingness as unimportant

E. Choosing Counts and Percentages

Counts indicate the absolute number of observations, whereas percentages enable comparisons across groups of different sizes. Both measures are generally necessary. Row percentages show how outcomes are distributed within each row group, while column percentages show how row categories are distributed within each column group. The selected approach should reflect the research question and be reported explicitly (Mishra et al., 2019).

Percentages based on very small denominators should not be presented without appropriate caution. In sparse tables, exact counts and confidence intervals may be more informative than percentages that imply unjustified precision (Vandenbroucke et al., 2014).

F. Missing Data and Unknown Categories

Missingness is an analytical finding rather than a formatting inconvenience. Tables that silently exclude missing observations may overstate data completeness and produce inconsistent denominators across variables. “Missing,” “unknown,” “refused,” and “not applicable” responses should be distinguished when they represent different circumstances (Little et al., 2012).

Researchers should report the number included in each analysis and assess whether missingness varies across subgroups. Complete-case tabulation may suit simple descriptions, but inferential analyses may require multiple imputation, weighting, or sensitivity analysis, depending on the missing-data mechanism (Sterne et al., 2009).

G. Classification and Table Quality Audit

Table 3. Classification-Tabulation-Interpretation Quality Framework

Quality principle	Audit question	Minimum evidence	Corrective action
Conceptual validity	Do categories represent the intended construct?	Operational definition linked to the research question	Revise definitions or collect a more valid measure
Mutual exclusivity	Can one observation enter more than one category unintentionally?	Coding rule and duplicate-category check	Redesign categories or use explicit multi-response coding
Exhaustiveness	Can every valid observation be coded?	Documented missing, other and not-applicable rules	Add categories or refine skip logic
Information preservation	Has grouping removed clinically or scientifically important variation?	Original raw values retained and cut-points justified	Analyse continuous values and provide sensitivity analysis
Denominator transparency	Is the percentage or rate denominator clear and appropriate?	Counts, totals and footnotes shown	Recalculate using a matched population or person-time denominator
Missing-data visibility	Are excluded or missing observations reported?	Missing counts by variable and subgroup	Add completeness table and assess missingness
Reproducibility	Can another analyst recreate the table?	Codebook, syntax and versioned dataset	Document transformations and save analysis scripts
Ethical disaggregation	Does subgroup reporting reveal inequity without risking identification?	Justified categories and disclosure controls	Aggregate sensitive cells or apply controlled access

IV. DISCUSSION AND APPLIED IMPLICATIONS

A. Classification Is an Analytical Decision

Classification influences every subsequent stage of analysis. A variable coded as nominal is summarised differently from the same construct measured as ordinal or continuous (Mishra et al., 2019). Once observations are collapsed into broad categories, lost information cannot be recovered unless the original values are retained (Altman & Royston, 2006). Researchers should therefore prespecify classification rules and preserve raw data separately from derived analytical variables to support transparency and reproducibility (Wilkinson et al., 2016).

B. Tables Are Diagnostic Tools

Tables do more than communicate findings. Before modelling, they can reveal impossible values, sparse categories, group imbalances, unexpected missingness, and potential confounding. A well-designed baseline table may identify differences between comparison groups, while a contingency table can indicate whether an apparent association is concentrated within a particular subgroup (Penev et al., 2024).

However, tables alone should not be interpreted as causal evidence. Observed differences may result from sampling, confounding, measurement error, or random variation. Tabulation therefore provides a foundation for further analysis rather than a substitute for appropriate study design and causal reasoning (Hernán & Robins, 2020).

C. Public-Health and Policy Implications

In health systems, classification rules influence surveillance counts, disease trends, and resource allocation. Changes in case definitions may produce apparent increases or decreases in reported incidence without corresponding changes in underlying disease occurrence (Ghildayal et al., 2024). Similarly, combining small geographical or demographic groups may conceal health inequities, whereas excessively detailed disaggregation may increase confidentiality and disclosure risks.

Policy tables should clearly report definitions, data sources, reference periods, denominators, and relevant measures of uncertainty (Vandenbroucke et al., 2014). Decision-makers must also be able to distinguish accurately among counts, proportions, rates, and modelled estimates to prevent inappropriate interpretation and policy responses (Janssen et al., 2024).

D. Software, Automation and Human Judgment

Software can automate coding, pivot tables, summaries, and validation, but cannot determine whether categories are scientifically meaningful (Penev et al., 2024). Automated workflows should incorporate range checks, label verification, duplicate detection, denominator validation, and reproducible scripts. Human oversight remains essential for assessing conceptual validity and ensuring responsible interpretation (Wilkinson et al., 2016).

E. FAIR and Reproducible Data Practice

Effective classification and tabulation support the FAIR principles, which require data to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Standardised variable names, controlled vocabularies, explicit measurement units, comprehensive data dictionaries, and version-controlled code enhance cross-study comparability and enable independent reproduction of statistical analyses (Finster et al., 2025).

F. Common Reporting Errors

i. Using percentages without reporting the underlying counts.
ii. Changing denominators across columns without explanation.
iii. Combining missing values with valid categories.
iv. Reporting too many decimal places and implying false precision.
v. Creating overlapping or non-exhaustive categories.
vi. Using unequal class intervals without adjusting interpretation.
vii. Presenting only statistically significant subgroup tables.
viii. Failing to state whether percentages are row, column or total percentages.
ix. Suppressing small cells inconsistently or without disclosure rules.
x. Using colour or visual emphasis that is inaccessible or analytically misleading.

V. CONCLUSION AND RECOMMENDATIONS

Data classification and tabulation are central components of statistical validity. Classification determines what a variable means and which analyses are defensible; tabulation determines how distributions, comparisons, missingness and possible errors become visible. When these tasks are performed poorly, technically correct calculations can still support misleading conclusions.

The strongest practice begins with clear construct definitions, preserves original information, uses mutually exclusive and exhaustive categories, reports denominators and missingness, and documents every transformation. Tables should be designed around specific research questions and should function as both communication tools and analytical quality checks.

Recommendations

i. Develop a data dictionary before analysis, including variable definitions, labels, units, coding rules and allowable values.
ii. Preserve raw continuous values and create grouped variables only as documented derivatives.
iii. Pre-specify category cut-points using scientific, clinical or policy rationale rather than searching for statistically convenient divisions.
iv. Display counts with percentages and identify clearly whether percentages are calculated by row, column or total.
v. Report missing values, exclusions and variable-specific sample sizes in every relevant table.
vi. Use reproducible syntax or scripts rather than undocumented manual recoding and table editing.
vii. Validate all tables through independent checks of totals, denominators, ranges and internal consistency.
viii. Apply disclosure controls when disaggregating small or sensitive subgroups.
ix. Train students and researchers to treat table design as statistical reasoning rather than cosmetic formatting.
x. Adopt metadata and interoperability standards to support comparison and reuse across datasets and institutions.

VI. DECLARATIONS

Ethics Approval and Consent to Participate

This methodological review used published literature and illustrative examples and did not involve direct participation by human subjects.

Funding

The author received no specific funding for this work.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

No primary dataset was generated. All methodological evidence is available from the cited sources.

Author Contribution

Dr. Daniel A. Otwori conceptualised the manuscript, conducted the methodological synthesis, developed the quality framework and prepared the final paper.

AI and Digital Tool Disclosure

Digital tools were used for language refinement, document formatting and reference organisation. The author reviewed the manuscript and accepts responsibility for its accuracy, interpretation and originality.

VII. REFERENCES

Agresti, A. (2013). Categorical data analysis (3rd ed.). Wiley.

Bluman, A. G. (2018). Elementary statistics: A step-by-step approach (10th ed.). McGraw-Hill Education.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). Sage.

Field, A. (2018). Discovering statistics using IBM SPSS Statistics (5th ed.). Sage.

Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten (2nd ed.). Analytics Press.

International Organization for Standardization. (2016). ISO 8000-61: Data quality - Part 61: Data quality management: Process reference model. ISO.

Kirkwood, B. R., & Sterne, J. A. C. (2003). Essential medical statistics (2nd ed.). Blackwell Science.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). Introduction to the practice of statistics (9th ed.). W. H. Freeman.

Rothman, K. J., Greenland, S., & Lash, T. L. (2021). Modern epidemiology (4th ed.). Wolters Kluwer.

Triola, M. F. (2018). Elementary statistics (13th ed.). Pearson.

United Nations Economic Commission for Europe. (2023). Generic Statistical Information Model: Reference framework for statistical information. UNECE.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

World Health Organization. (2017). Data quality review: A toolkit for facility data quality assessment. WHO.

World Health Organization. (2021). Guidance for health information system data quality and use. WHO.

VIII. COPYRIGHT AND LICENSE