Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Methodological alignment and reporting in secondary analyses of DHS, MICS and DHIS2 data: a systematic meta-research study in the East African Community

TUYISHIME AUDRE PRINCE

priplee@gmail.com

University of Rwanda

Patrick Kayinamura Muhirwa

Rwanda Biomedical Center

Research Article

Keywords: complex surveys, data quality, DHIS2, methodological study, meta-research, routine health information systems, secondary data, survey methods

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11499274/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Methodological alignment and reporting in secondary analyses of DHS, MICS and DHIS2 data: a systematic meta-research study in the East African Community

Authors: TUYISHIME AUDRE PRINCE[1*]; Patrick Kayinamura Muhirwa[2]

Affiliations:

[1] School of Public Health, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda.

[2] One Health Unit, Division of Public Health Surveillance and Emergency Preparedness and Response, Rwanda Biomedical Centre, Kigali, Rwanda.

Corresponding author: TUYISHIME AUDRE PRINCE, School of Public Health, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. Email: priplee@gmail.com. ORCID: 0009-0002-0799-3140.

Co-author ORCID: Patrick Kayinamura Muhirwa, 0009-0005-9629-350X.

Abstract

Background: Secondary analyses of complex surveys and routine health information systems require source-specific methodological safeguards, but published reports do not always distinguish documented analytical incompatibility from incomplete reporting. We estimated how

often data-generating structure, inferential target, statistical method and interpretation were aligned in contemporary East African analyses of Demographic and Health Survey (DHS), Multiple Indicator Cluster Survey (MICS) and DHIS2 data.

Methods: We conducted a systematic methodological study of peer-reviewed single-country analyses published from January 2020 through August 2026. The identified eligible frame contained 466 publications (377 DHS/MICS and 89 DHIS2). A stratified probability sample of 127 publications was coded across four linked domains as aligned, materially misaligned, insufficiently reported or not applicable, with finite-population inference; 26 publications were independently coded for reliability. A prespecified Rwanda illustration compared full survey-design, unweighted and weights-only analyses.

Results: Fully aligned publications were estimated at 38.7% for DHS/MICS and 40.7% for DHIS2. Material misalignment was 17.3% and 39.2%, while insufficient reporting was 44.0% and 20.1%, respectively. The proportion not fully aligned was 61.3% for DHS/MICS and 59.3% for DHIS2 (difference -2.0 percentage points, 95% confidence interval -15.6 to 11.6). Among not fully aligned publications, documented material misalignment accounted for 28.2% of the DHS/MICS total and 66.1% of the DHIS2 total. Overall pre-adjudication agreement was 84.6% (Cohen kappa 0.733; Gwet AC1 0.785). In the 2025 Rwanda DHS, unweighted and weights-only prevalence standard errors were 61.8% and 67.5% of the full-design standard error; the higher-education odds ratio was 8.34 unweighted versus 6.33 with the full survey design.

Conclusions: Similar overall proportions not fully aligned masked different methodological problems: insufficient observability predominated in DHS/MICS, whereas documented incompatibility contributed more in DHIS2. Methodological assessment and reporting should separate demonstrated misalignment from missing information and apply source-specific safeguards for complex survey design and routine-data provenance, denominators, completeness and temporal structure.

Keywords: complex surveys; data quality; DHIS2; methodological study; meta-research; routine health information systems; secondary data; survey methods

Background

Methodological studies evaluate the design, analysis or reporting of other research and can identify threats to the validity and interpretability of published evidence [1]. Secondary health data are a particularly important setting because the same statistical procedure can be appropriate or inappropriate depending on how observations were generated and what quantity is being estimated. In the East African Community (EAC), two data ecosystems are especially influential. Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) are designed probability samples intended to support population inference, whereas routine health information systems, commonly implemented through DHIS2, record service and surveillance data generated through health-system processes. Valid secondary analysis therefore depends on alignment between the data-generating structure, inferential target, statistical method, uncertainty calculation and interpretation.

Concern about secondary analysis of complex surveys is longstanding. Bell and colleagues reviewed 1,003 empirical articles using three large health surveys and found that only 60% reported accounting for design effects and 61% reported using sample weights [2]. West, Sakshaug and Aurelien subsequently examined analytic errors in secondary analyses of national survey data and showed that incorrect handling of complex designs can materially alter estimates and standard errors [3]. These studies establish an important scholarly lineage for the present work. Our contribution is not to rediscover that complex-survey analysis can be mishandled, but to distinguish positively documented incompatibility from insufficient reporting and to place survey and routine-data analyses within the same structure-to-target-to-method-to-interpretation framework.

For DHS analyses, official guidance emphasises sampling weights, clusters or primary sampling units, strata and design-appropriate variance estimation [4]. Methodological work using Tanzanian survey data has likewise shown that weighting and variance choices can alter bias, variance and confidence-interval performance [5]. A multilevel model can address hierarchical structure, but it does not automatically substitute for design-based survey inference. The relevant question is whether the method matches the estimand and the features of the survey design that matter for that estimand.

Routine facility data require different safeguards. World Health Organization guidance for routine health information systems highlights representativeness, completeness and timeliness, internal consistency, population denominators and interpretation as core analytical concerns [6,

7]. Empirical assessments across Eastern and Southern Africa have documented outliers, temporal inconsistencies and denominator problems even where reporting completeness is high [8]. DHIS2 methodological work has similarly emphasised missing-versus-zero ambiguity, unstable or poorly documented denominators, heterogeneous indicator quality and the need for explicit cleaning and validation workflows [9]. Related routine-data literature has addressed evaluation design [10], missing-data handling [11] and strengthening information systems for analysis and data use [12]. These are not peripheral data-management issues: they can change the meaning of a rate, trend, association or intervention contrast.

Previous survey-methods and routine-data literatures have largely developed in parallel. A cross-source comparison therefore needs a framework that respects source-specific structures without treating the domains as interchangeable or statistically independent. We used four linked diagnostic domains: D1, data-generating structure; D2, inferential goal; D3, statistical method; and D4, reporting or interpretation. These domains form a diagnostic chain rather than four independent latent traits.

We conducted a systematic methodological study and meta-research analysis of contemporary single-country EAC studies using DHS, MICS or DHIS2 data. Our primary objective was to estimate, within an identified eligible publication frame, the frequency and pattern of full alignment, positively documented material misalignment and insufficient reporting. We also examined prespecified source-specific sentinel problems, assessed inter-reviewer reliability and used an independently frozen Rwanda survey analysis to illustrate the inferential consequences of one recurrent review finding.

Methods

Study design, protocol, and registration

We conducted a systematic methodological study of published secondary-data analyses, with the article as the unit of analysis. The design combined a systematic review frame with probability-based methodological coding and finite-population inference. It was not a pooled analysis of published effect estimates or a bibliometric analysis. The protocol prespecified separate primary reporting for designed household-survey analyses and routine-health-information analyses and defined the four linked domains D1 to D4.

The protocol, eligibility rules, coding framework, sampling strategy and reviewer instructions were frozen before production methodological outcome coding. The Open Science Framework registration was made public on 20 August 2026 [13]. Search development and execution, candidate-frame construction, calibration and instrument freeze had occurred before registration. Main title-abstract eligibility decisions, full-text eligibility decisions, detailed methodological outcome coding, publication-sample selection and the Rwanda consequence analysis had not begun. This prospective boundary is retained in the audit trail. The review is reported in accordance with PRISMA 2020 where applicable to this systematic methodological study [14]; a completed checklist is provided in Supplementary Table S8 (Additional file 1).

Eligibility

Eligible articles were peer-reviewed full original research articles published from 1 January 2020 through the final August 2026 search, with a primary single-country analysis from Burundi, the Democratic Republic of the Congo, Kenya, Rwanda, Somalia, South Sudan, Tanzania, or Uganda. For DHS/MICS, eligible studies quantitatively analysed individual- or household-level microdata, and every survey round contributing to the primary analysis had to be fielded in 2015 or later. For DHIS2, eligible studies quantitatively analysed routinely collected health-service, surveillance, or programme data held, managed, synchronised, or aggregated in DHIS2, with at least one observation from 2015 onward. Data-quality studies were eligible.

Searches were not restricted by language. At full-text eligibility, however, an official complete English methodological text was required for assessment; an English abstract alone was insufficient. Records without an official complete English full text were excluded as “full text not available in English”, and translation was not used to make a record eligible.

Implementation-only papers, qualitative-only studies, reviews, protocols, non-separable mixed-source analyses, aggregate-only DHS/MICS analyses, and duplicate or companion reports were excluded under a frozen hierarchical reason system. Analyses had to be single-country; multi-country pooled analyses were excluded from the primary review, as specified in the frozen protocol. Methodological weakness was not itself an eligibility exclusion. If an article was otherwise eligible, incomplete reporting of a specific analytical feature was represented later in the A/M/I framework rather than used to remove the article.

Search, deduplication, and screening

Separate search families were used for DHS/MICS and DHIS2/RHIS. The final PubMed production run on 19 August 2026 comprised 18 queries and returned 603 PMIDs representing 602 publications after one duplicate-index adjudication. Scopus production retrieval was performed through Publish or Perish 8.19.5300.9483 using 18 frozen RIS exports. These contained 848 raw records and 793 unique Scopus publications after exact-record deduplication. PubMed and Scopus shared 556 publication-level records, giving a primary bibliographic union of 839 publications. The bibliographic sources were prespecified as PubMed and Scopus, with the DHS Program journal-article database used as a mandatory supplementary discovery source for the survey arm. African Journals Online, Global Health and Web of Science were not searched; anchor-recall testing and supplementary discovery were used to audit the implemented strategy, but they do not establish complete literature capture.

The prespecified DHS Program supplementary search identified 159 publication-level records, of which 68 were already in the primary union and 91 were additional. The final frozen candidate frame therefore contained 930 publications. Search queries, raw counts, export hashes, anchor checks, deduplication adjudications, and the candidate-frame SHA-256 are retained in the reproducibility record (Supplementary Table S7 in Additional file 1 and public search-audit companion files). Search-anchor checks were used as regression tests for known indexed records and were not interpreted as estimates of literature-search sensitivity.

ASReview LAB 3.0.8 was used only to prioritize screening order. No model-based stopping rule or automatic eligibility decision was used, and all 930 records received a human title-abstract

decision. Twenty-seven records remained title-only after metadata enrichment and were protected by a safeguard that prevented definitive exclusion without full-text assessment. A quality-control re-audit of the 93 initially excluded title-abstract records restored 24 to full-text review, leaving 69 final title-abstract exclusions and 861 records in the full-text frame.

Full-text eligibility was protocol driven. At the independent-review gate, R1 and R2 assessed the 820 records for which complete screening evidence was then available, and disagreements were adjudicated against the frozen eligibility rules and article evidence. The remaining 41 records were resolved after sufficient evidence became available. Final exclusion reasons were harmonised only after eligibility verdicts were locked, using the prespecified first-applicable hierarchy (Supplementary Table S2 in Additional file 1). No eligibility verdict was changed by reason harmonization.

Finite publication frame and stratified probability sample

The final identified eligible frame contained 466 publications: 377 DHS/MICS, comprising 370 DHS and 7 MICS publications, and 89 DHIS2 publications. These 466 publications define the finite population for the primary review estimands. The estimands are conditional on the frozen search and eligibility frame and are not presented as superpopulation prevalences for all publications that might exist outside that frame.

Because detailed coding of all 466 publications was not required by the prespecified precision target, we selected a stratified sample without replacement by data family and country. The

frozen sample contained 80 DHS/MICS and 47 DHIS2 publications, total n=127. Allocation was near proportional with sparse-country protection. Selection was instantiated reproducibly within each stratum by ranking the SHA-256 value of the frozen seed, data family, country, and publication ID. Publication inclusion probabilities and inverse-probability sampling weights were retained (Supplementary Table S1 in Additional file 1). The planned worst-case arm-level 95% confidence-interval half-widths at p=0.50 were approximately 9.7 percentage points for DHS/MICS and 10.0 percentage points for DHIS2.

One originally selected Kenya DHS publication could not be completely coded because complete accessible article text was unavailable at the production-coding stage. The analysis dataset nevertheless contains 127 coded publications because the next deterministic record in the same stratum was used as a reserve; the inaccessible originally selected publication itself has no coded outcome. For the primary analysis, the Kenya DHS stratum retains the frozen $N_h=64$ and $n_h=14$ weighting as a coding-nonresponse adjustment. A sensitivity analysis instead treats the codeable frame as $N_h=63$ with $n_h=14$. No outcome was imputed for the inaccessible publication.

Five pilot-exposed selected publications remained in the target population but were recoded de novo under the frozen production instrument.

Coding framework and outcomes

Each sampled article was coded across four linked domains: D1, data-generating structure; D2, inferential goal; D3, choice of method; and D4, reporting and interpretation. These domains were designed as a diagnostic chain from structure to target to method to interpretation. They were not treated as four statistically independent observations or as four interchangeable measurements of a single latent quality construct.

Each applicable domain was classified as aligned (A), materially misaligned (M), insufficiently reported to determine alignment (I), or not applicable (NA). An M classification required a positively reported or demonstrated incompatibility capable of materially changing the target, estimate, uncertainty, or substantive interpretation. Silence or incomplete reporting was coded I rather than assumed to be error.

Overall classification was rule based. An article was materially misaligned if any applicable core domain was M; fully aligned if all applicable domains were A; and indeterminate or partial if no domain was M but at least one applicable domain was I. No additive quality score was created.

DHS/MICS-specific fields covered weights, clustering, stratification, analytic domains, variance estimation, multilevel alignment, and survey-design reporting. DHIS2-specific fields covered platform provenance, reporting completeness and timeliness, missing-versus-zero handling, numerator and denominator provenance, numerator-denominator alignment, implausible values, cleaning, temporal structure, and reporting-unit dependence.

Twelve prespecified sentinel problems were coded only when applicable and observable. Confirmed adverse classifications required an evidence location and short factual basis. Sentinel prevalence was always interpreted relative to its applicable domain; small applicable denominators were not elevated into population-wide conclusions.

Reviewers and reliability

R1 coded all 127 sampled publications. R2 independently coded a frozen reliability subsample of 26 publications, 16 DHS/MICS and 10 DHIS2, without access to R1 classifications. Reliability was calculated from preserved independent pre-adjudication classifications. Disagreements were then adjudicated before the final analysis dataset was frozen.

Raw agreement, Cohen kappa, and Gwet AC1 [15] were reported separately for D1, D2, D3, D4, and the rule-based overall classification (Supplementary Table S3 in Additional file 1). No pooled D1-D4 kappa was used because the four domain judgments belong to the same publication and are not independent repeated observations.

Statistical analysis

Primary arm-level estimands were finite-population proportions among the 377 identified eligible DHS/MICS publications and 89 identified eligible DHIS2 publications. Estimates used the frozen stratified publication-sampling design, inverse inclusion-probability weights, and finite-population correction. Variance estimation used the survey package in R with the

publication stratum as the sampling stratum. Proportion intervals used Taylor-linearized standard errors with t critical values based on design degrees of freedom and were bounded to [0,1]. Between-arm contrasts used independent-arm variance sums with Welch-Satterthwaite degrees of freedom. The primary lonely-stratum rule was survey.lonely.psu = "adjust"; alternative lonely-stratum rules were examined as sensitivity analyses.

Ninety-five per cent confidence intervals quantify uncertainty introduced by subsampling publications for detailed coding. They do not quantify uncertainty from incomplete literature discovery, eligibility misclassification, reviewer classification error, or a hypothetical superpopulation of future publications.

The primary three-state outcome retained A, M, and I separately. Because M is an observed documented incompatibility and I is an observability state, M was not interpreted as the latent error probability among I publications. We also estimated the combined proportion not fully aligned, M+I, and descriptively decomposed that total into the share attributable to documented M versus insufficiently reported I. This composition analysis is an observability analysis, not a model for hidden error; the M/(M+I) decomposition is descriptive and is not assigned a separate null-hypothesis or equivalence test.

For D1-D4, arm-specific state proportions were estimated separately. Six pairwise cross-tabulations described how domain states co-occurred within publications (Supplementary Table S6 in Additional file 1). No chi-square test, pooled-domain variance calculation, or independence claim was made from those cross-tabs.

For sentinel problems and other fields not applicable to all publications, prevalence was estimated within the applicable publication domain using a ratio of Horvitz-Thompson totals with Taylor linearization. Confirmed adverse and unclear states remained separate. Between-arm differences in primary proportions used the independent-arm variance sum.

Sensitivity analyses compared weighted with unweighted estimates, excluded the five pilot-exposed publications, estimated material misalignment among determinate articles only, combined M and I as not fully aligned, varied the lonely-stratum variance rule, and redefined the Kenya DHS codeable frame from 64 to 63 publications (Supplementary Table S5 in Additional file 1). All literature-review estimates and figures were generated in R 4.6.0 using survey 4.5 and version-controlled analysis code.

Prespecified empirical illustration of inferential consequences

A separate Rwanda module was used as an empirical illustration of inferential consequences, not as validation of the literature classifications. The scenarios were selected from review findings and frozen before numerical analysis of the 2025 Rwanda DHS microdata. The 2019-20 RDHS [16] was used as a benchmark and the 2025 RDHS [17] as the contemporary illustration. The official 2025 Final Report and survey metadata document a two-stage stratified design with 560 selected clusters and 60 sampling strata [17].

For a weighted health-insurance prevalence and for an education-association logistic regression, we compared three analyses: a full design-based analysis incorporating sampling weights, clusters, and strata; an unweighted analysis treating records as independent; and a weights-only analysis that used the survey weights but omitted the cluster and stratum structure. The purpose was to quantify how much point estimates and standard errors changed under realistic simplifications. The module was not used to infer the correctness of any reviewed article and no threshold for a desirable effect size was specified after seeing the results.

Reproducibility and research ethics

Search logic, search execution manifests, export hashes, deduplication adjudications, candidate-frame hashes, eligibility rules, sampling seeds, inclusion probabilities, coding dictionaries, derived review data, R source code, session information and generated outputs are version controlled. The private research repository retains operational audit material. A public reproducibility repository [18] contains release-safe code, derived data, metadata and figures; copyrighted full texts, licensed database exports and restricted DHS microdata are not redistributed.

No new human-participant data were collected for the literature audit. The Rwanda empirical illustration used anonymised survey microdata obtained through authorised access from The DHS Program [19] under the applicable data-use conditions.

Use of generative AI tools

OpenAI ChatGPT (GPT-5.6 Sol) was used for code drafting, data checks, literature organisation and language editing. The authors reviewed and verified all outputs and made all screening, eligibility, coding and adjudication decisions.

Results

Search and study flow

The frozen PubMed search contained 602 publication-level records and the deduplicated Scopus search contained 793. Their overlap was 556 publications, producing a primary bibliographic union of 839. The DHS Program supplementary search contributed 91 additional unique publications, giving 930 candidate records.

After title-abstract quality control, 69 records were excluded before full-text review and 861 entered the full-text frame. Full-text review excluded 395 and left 466 eligible publications. The largest full-text exclusion categories were wrong geography or multi-country primary analysis (n=163), use of DHS/MICS rounds before 2015 (n=139), wrong data family or source (n=28), and not being an original full research article (n=21).

The eligible finite frame comprised 377 DHS/MICS publications and 89 DHIS2 publications. The detailed probability sample contained 80 DHS/MICS and 47 DHIS2 publications. The

reliability subsample contained 26 publications. The identification and screening flow is shown in Supplementary Figure S2 (Additional file 2).

Sampling diagnostics

The final publication-sampling design remained close to proportional allocation. The coefficient of variation of the final weights was 0.046 for DHS/MICS and 0.114 for DHIS2, corresponding to unequal-weight design effects of 1.002 and 1.013. Weighted and unweighted primary point estimates differed by no more than 1.2 percentage points.

Overall alignment and observability

The estimated proportion of fully aligned publications had close point estimates: 38.7% in DHS/MICS (95% CI 29.1% to 48.3%) and 40.7% in DHIS2 (30.9% to 50.5%). The estimated difference, DHIS2 minus DHS/MICS, was 2.0 percentage points (95% CI -11.6 to 15.6).

The complementary estimates of publications not fully aligned were 61.3% in DHS/MICS and 59.3% in DHIS2. Their difference, DHIS2 minus DHS/MICS, was -2.0 percentage points (95% CI -15.6 to 11.6). That interval is compatible with non-trivial differences in either direction, so the primary totals are not interpreted as an equivalence result. The descriptive composition of non-full classifications showed a more marked contrast. In DHS/MICS, 17.3% of all eligible publications were positively documented as materially misaligned and 44.0% were indeterminate or partially reported. In DHIS2, the corresponding estimates were 39.2% and 20.1%. Table 1

summarises these estimates, and Figure 1 shows the corresponding overall estimates with 95% confidence intervals.

Table 1. Primary overall alignment estimates by data-source family.

Source	Overall state	Raw n/N	Weighted %	95% CI
DHS/MICS	Fully aligned	31/80	38.7	29.1 to 48.3
DHS/MICS	Materially misaligned	14/80	17.3	9.7 to 25.0
DHS/MICS	Indeterminate / partial	35/80	44.0	34.2 to 53.8
DHIS2	Fully aligned	19/47	40.7	30.9 to 50.5
DHIS2	Materially misaligned	19/47	39.2	30.0 to 48.4
DHIS2	Indeterminate / partial	9/47	20.1	11.8 to 28.4

Among publications classified as not fully aligned, documented M therefore accounted for an estimated 28.2% of the DHS/MICS total and 66.1% of the DHIS2 total; insufficiently reported I accounted for 71.8% and 33.9%, respectively. These quantities describe the observed composition of non-alignment and do not assign hidden error status to I publications.

Core-domain pattern and domain coupling

D2, inferential goal, was aligned in approximately 98% of both data-source families. The larger differences occurred in D1, D3, and D4. For DHS/MICS, D1 was aligned in 38.7%, materially misaligned in 15.1%, and insufficiently reported in 46.2%. D3 was aligned in 39.9%, materially misaligned in 17.3%, and insufficiently reported in 42.8%.

For DHIS2, D1 was aligned in 58.0%, materially misaligned in 17.4% and insufficiently reported in 24.6%. D3 was aligned in 45.1%, materially misaligned in 39.2% and insufficiently reported in 15.7%. D4 was aligned in 95.3% of DHS/MICS publications, with 3.4% materially misaligned and 1.3% insufficiently reported. In DHIS2, D4 was aligned in 69.6% and materially misaligned in 30.4%. Table 2 summarises the domain-specific estimates, and Figure 2 shows the corresponding alignment-state distributions.

Table 2. Core-domain alignment states by data-source family.

Source	Domain	A: raw n (weighted %)	M: raw n (weighted %)	I: raw n (weighted %)
DHS/MICS	D1	31 (38.7%)	12 (15.1%)	37 (46.2%)
DHS/MICS	D2	78 (97.8%)	2 (2.2%)	0 (0.0%)
DHS/MICS	D3	32 (39.9%)	14 (17.3%)	34 (42.8%)
DHS/MICS	D4	76 (95.3%)	3 (3.4%)	1 (1.3%)
DHIS2	D1	28 (58.0%)	8 (17.4%)	11 (24.6%)
DHIS2	D2	46 (97.9%)	1 (2.1%)	0 (0.0%)
DHIS2	D3	21 (45.1%)	19 (39.2%)	7 (15.7%)
DHIS2	D4	32 (69.6%)	15 (30.4%)	0 (0.0%)

Because D1-D4 are linked judgments within the same publication, domain-state cross-tabulations were treated descriptively rather than as independent observations. The complete pairwise cross-tabs are reported in the supplement.

Source-specific structural findings and sentinel problems

For DHS/MICS, confirmed misuse of sampling weights was uncommon, estimated at 2.5%, but uncertainty about design handling was much more frequent. Survey variance handling was

positively documented as inappropriate in 12.6% and remained unclear in 46.2%. Cluster or PSU handling was unclear in 58.8%, and stratification handling was unclear in 61.3%. Only 23.6% of DHS/MICS publications had survey-design reporting classified as sufficient.

For DHIS2, reporting completeness was explicitly not considered in an estimated 6.5% of publications and remained unclear in 47.6%. Explicitly inappropriate missing-versus-zero handling was estimated at 4.5%. Among publications for which denominators were applicable, denominator provenance was insufficient in 15.0% and numerator-denominator alignment was not considered in 15.0%. Routine temporal structure was materially mishandled in 24.7% of applicable publications.

Sentinel findings were retained as supporting diagnostics rather than ranked as a second primary outcome. In DHS/MICS, the broad complex-survey sentinel was present in 16.0% of applicable publications. In DHIS2, the broad routine-indicator data-quality or denominator sentinel was present in 33.8%, and reporting or interpretation overclaim was present in 30.4%. Full estimates for all 12 sentinels, including applicable denominators and unclear states, are reported in Supplementary Table S4 (Additional file 1); graphical prevalence estimates are shown in Supplementary Figure S1 (Additional file 3). Sentinels with very small applicable domains are not emphasised in the main text.

Sensitivity analyses

Weighting had little influence because allocation was close to proportional. The maximum absolute weighted-versus-unweighted difference among the primary A/M/I outcomes was 1.2 percentage points. Excluding pilot-exposed publications changed each primary arm estimate by less than 2 percentage points; material misalignment was then estimated at 17.9% in DHS/MICS and 38.5% in DHIS2.

Among publications with determinate overall classifications only, material misalignment was estimated at 30.9% in DHS/MICS (95% CI 18.5% to 43.3%) and 49.1% in DHIS2 (38.3% to 59.8%). Alternative lonely-stratum rules affected uncertainty but not point estimates. The Kenya coding-nonresponse sensitivity is reported with the regenerated R outputs and is interpreted against the frozen-frame primary analysis.

Inter-reviewer reliability

Twenty-six publications were independently coded before adjudication. Agreement for the overall alignment classification was 84.6%, with Cohen kappa 0.733 and Gwet AC1 0.785. Domain-specific raw agreement was 80.8% for D1, 100% for D2, 80.8% for D3 and 92.3% for D4. Corresponding kappas were 0.654 for D1, not estimable for D2 because both reviewers classified all 26 publications as aligned, 0.669 for D3 and 0.469 for D4; corresponding AC1 values were 0.735, not estimable, 0.732 and 0.917. The D4 contrast illustrates the effect of marginal imbalance on kappa. Among the 13 publications that both reviewers classified as not fully aligned, M-versus-I agreement was 84.6% (11/13), with kappa 0.435 and AC1 0.792. This small conditional subset is reported descriptively because the M-versus-I distinction is central to

the paper. No pooled D1-D4 kappa was calculated or interpreted. Table 3 summarises the pre-adjudication reliability results.

Table 3. Pre-adjudication inter-reviewer reliability.

Measure	Agreement n/N	Raw agreement	Cohen κ	Gwet AC1
D1	21/26	80.8%	0.654	0.735
D2	26/26	100.0%	NE	NE
D3	21/26	80.8%	0.669	0.732
D4	24/26	92.3%	0.469	0.917
Overall	22/26	84.6%	0.733	0.785
M vs I, jointly non-full	11/13	84.6%	0.435	0.792

Prespecified empirical illustration of inferential consequences

In the 2019-20 RDHS benchmark, the full-design estimate of health-insurance coverage was 82.65%, with a standard error of 0.549 percentage points. Treating observations as independent produced a standard error only 56.2% as large as the full-design standard error, corresponding to a full-design versus simple-random-sample design effect (DEFF) of 3.16 and design factor (DEFT) of 1.78. The weights-only standard error was 61.5% as large as the full-design value; the corresponding full-design versus weights-only variance ratio was 2.65.

In the 2025 RDHS, full-design health-insurance coverage was 87.43%, with a standard error of 0.442 percentage points. The unweighted point estimate was 87.90%, 0.47 percentage points higher, and its standard error was only 61.8% as large as the full-design standard error, corresponding to DEFF 2.62 and DEFT 1.62. The weights-only analysis recovered the same

prevalence point estimate as the full-design analysis but produced a standard error only 67.5% as large; the full-design versus weights-only variance ratio was 2.20.

For the education association in 2025, the full-design odds ratio for higher versus no education was 6.33, compared with 8.34 in the unweighted analysis. Across the prespecified contrasts, simplification changed precision and effect magnitude but did not reverse an association or change whether the confidence interval included one. The empirical illustration therefore supports a narrow consequence claim: omitting design structure can materially overstate precision and can shift effect magnitude, even when broad substantive direction is unchanged. Figure 3 shows the corresponding standard-error ratios.

Discussion

The central methodological finding is not that one data source was worse than the other. In the prespecified probability-sample analysis, the proportions not fully aligned were close, 61.3% for DHS/MICS and 59.3% for DHIS2, but the DHIS2-minus-DHS/MICS difference was imprecisely estimated at -2.0 percentage points (95% CI -15.6 to 11.6). We therefore do not interpret the primary totals as evidence of equivalence. The more informative result was the observable composition of non-alignment: DHS/MICS publications more often lacked enough methodological information to establish alignment, whereas DHIS2 publications more often contained positively documented incompatibilities. Separating demonstrated incompatibility from insufficient observability is the main interpretive contribution of the A/M/I framework and avoids treating missing reporting as latent analytical error.

The survey-side result extends, rather than replaces, earlier meta-research. Bell and colleagues documented incomplete reporting of design effects and sample weights in a large body of secondary health-survey analyses [2]. West, Sakshaug and Aurelien showed that analytic errors in secondary survey use can be common and consequential [3]. Our DHS/MICS findings are consistent with that literature, but the explicit I state changes the interpretation. If a publication reports use of a weight but does not identify whether clustering, strata, analytic domains or design-based variance entered the analysis, the published record does not establish correct handling. It also does not establish that the analysis was wrong. Separating M from I prevents both over-accusation and false reassurance.

The DHIS2 pattern was different. World Health Organization routine-health-information guidance treats completeness, timeliness, internal consistency, denominator quality and interpretation as core components of analysis [6, 7]. Empirical assessments in Eastern and Southern Africa have documented outliers, temporal inconsistencies and implausible coverage values related to denominator problems [8]. Farnham and colleagues similarly described unreliable denominators, missing-versus-zero ambiguity and indicator heterogeneity as recurring hazards in DHIS2 analysis [9]. In our review, the higher M component in DHIS2 was concentrated in method choice and interpretation rather than in an inability to infer the research question itself.

The D1-D4 framework should be interpreted as a diagnostic chain. D1 asks what structure generated the observations, D2 what quantity or inferential target the article seeks, D3 whether the method can estimate that target from that structure, and D4 whether the reported interpretation remains within what the analysis supports. These are coupled judgments by design. The paper therefore does not use pooled-domain kappa, independence tests across domains, or an additive quality score. Domain-specific reliability and within-publication cross-tabs are more coherent with the framework.

The Rwanda module provides an empirical consequence check for one recurrent survey-side issue without pretending to validate the review classifications. The scenarios were selected from the review and frozen before the 2025 numerical outcomes were analysed. In the 2025 RDHS, the unweighted and weights-only analyses produced standard errors only 61.8% and 67.5% as large as the full-design standard error for health-insurance prevalence. The higher-education odds ratio shifted from 6.33 under the full design to 8.34 when analysed unweighted. These differences did not reverse the substantive direction, but they show why design reporting cannot be dismissed as cosmetic. The empirical module demonstrates consequence in an independent contemporary survey; it does not estimate how often the reviewed articles' conclusions would have changed.

Methodological implications for analysis and reporting

The findings support source-specific safeguards rather than a single generic quality rule. For DHS/MICS analyses, manuscripts should identify the survey weight, PSU or cluster variable, strata, analytic-domain handling where relevant and the variance-estimation approach. Multilevel

552 modelling should not be presented as automatic evidence that survey design was accommodated.
553 For DHIS2 analyses, provenance should be stated at the indicator and reporting-unit level, with
554 explicit treatment of completeness, missing versus zero, numerator and denominator provenance,
555 cleaning rules, temporal structure and reporting-unit dependence. When population incidence or
556 coverage is claimed, the denominator should represent the target population at risk. These
557 safeguards are practical implications of the observed failure modes and established source-
558 specific guidance, not a newly validated reporting checklist.
559
560 Several features strengthen the study. The candidate-frame construction and search execution are
561 auditable to query, export, hash, and deduplication level. The detailed review used a frozen
562 probability sample rather than a convenience corpus, and the estimands are explicitly tied to the
563 finite identified publication frame. The A/M/I rule separates demonstrated incompatibility from
564 insufficient observability. The production coding instrument, sampling design, reliability subset,
565 and primary analysis were frozen before the Rwanda consequence analysis. The numerical
566 review results are generated in R from an analysis-ready dataset rather than transcribed
567 manually.
568
569 The study also has limitations. First, the primary finite frame is conditional on the implemented
570 searches, supplementary discovery, deduplication, and eligibility process. It is an identified
571 bibliographic frame, not a census of all EAC scholarship, particularly work outside the indexed
572 and supplementary sources searched. The omission of African Journals Online, Global Health
573 and Web of Science may have excluded eligible publications, and that discovery uncertainty is
574 not represented by the publication-subsampling confidence intervals. The requirement for an

official complete English full text at eligibility may also have excluded otherwise eligible studies and introduced language-selection bias. The confidence intervals quantify publication-subsampling uncertainty within that frame; they do not quantify missed literature or justify inference to all possible future publications. Second, we evaluated what was reported in articles and accessible supplementary materials, not authors' private code or undocumented workflows. Some I classifications may therefore represent reporting omissions rather than incorrect analysis. Third, MICS contributed only seven eligible publications and two sampled publications, so it could not be analysed separately from DHS. Fourth, DHIS2 publications were heterogeneous in programme area, reporting unit, observation period, and target, and some sentinel applicable domains were small. Fifth, R1 coded the full detailed sample and R2 independently coded approximately one fifth; full duplicate coding might have reduced residual classification error. Chance-corrected reliability depended on the statistic and marginal distribution: D4 had 92.3% raw agreement, kappa 0.469 and AC1 0.917. The M-versus-I conditional check involved only 13 jointly non-full publications, so its 84.6% agreement should be interpreted with corresponding small-sample caution. Sixth, one selected Kenya DHS publication was a coding-stage nonresponse and was replaced by the next deterministic same-stratum reserve. This is handled transparently with a frozen-frame primary analysis and a codeable-frame sensitivity analysis. Seventh, complete accessible full text was not available for every eligible publication outside the primary detailed-coding sample. No abstract-only methodological outcome was imputed for those records. Because the primary estimates derive from the frozen probability sample, this access limitation does not alter the prespecified primary estimands. Finally, the Rwanda module examined two prespecified analytic scenarios in one country. It illustrates inferential

consequences and should not be generalized as a validation of every review classification or data source.

Conclusion

In this systematic methodological study, similar overall proportions of DHS/MICS and DHIS2 publications were not fully aligned, but the composition differed: insufficient reporting predominated in DHS/MICS, whereas documented incompatibility contributed more in DHIS2. Methodological assessment should therefore distinguish demonstrated misalignment from insufficient observability and should evaluate the full chain from data-generating structure to estimand, method, uncertainty and interpretation. Source-specific reporting and analytical safeguards are needed for both complex survey and routine health information system analyses.

List of abbreviations

A, aligned; AC1, Gwet's agreement coefficient 1; CI, confidence interval; DEFF, design effect; DEFT, design factor; DHS, Demographic and Health Survey; DHIS2, District Health Information Software 2; EAC, East African Community; I, insufficiently reported to determine alignment; M, materially misaligned; MICS, Multiple Indicator Cluster Survey; NA, not applicable; OSF, Open Science Framework; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PSU, primary sampling unit; RDHS, Rwanda Demographic and Health Survey; RHIS, routine health information system; R1, reviewer 1; R2, reviewer 2.

Declarations

Ethics approval and consent to participate

The literature-audit component did not involve human participants. The Rwanda empirical illustration used anonymised secondary microdata from the 2019-20 and 2025 Rwanda Demographic and Health Surveys obtained through authorised access from The DHS Program [19]. The protocols for both surveys were reviewed and approved by the Rwanda National Ethics Committee and the ICF Institutional Review Board [16, 17]. The original surveys obtained informed consent from respondents in accordance with their approved protocols. Because the present study was a secondary analysis of de-identified data, with no participant recruitment, contact or access to identifiable information, no further ethical approval or participant consent was required for this analysis.

Clinical trial number: not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Derived review datasets, coding dictionaries, domain cross-tabs, search-frame audit files, R source code, session information and aggregate Rwanda empirical-illustration outputs are available in the public reproducibility repository [18], frozen at commit 5f43ebbb08f9a8f02c883338d0caab5e3430f2b6. The analysis code was written in R 4.6.0, is released under the MIT License, and the repository includes session information needed to reproduce the release-safe analyses. Copyrighted article full texts and licensed database exports are not redistributed. The 2019-20 and 2025 RDHS respondent microdata are available from The

DHS Program to registered users under its data-use conditions [19] and are not redistributed by the authors.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

TAP conceptualised the study; developed the methodology; conducted the investigation, data curation, formal analysis and software work; performed validation and visualisation; drafted the manuscript; revised the manuscript; and administered the project. PKM contributed to investigation, independent validation and manuscript review and editing. Both authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Figure legends

Figure 1. Overall methodological alignment in the finite eligible publication frame. Design-weighted estimates and 95% confidence intervals for fully aligned, materially misaligned and indeterminate or partially reported publications, shown separately for DHS/MICS and DHIS2. Open and filled plotting symbols distinguish the two data-source families; colour does not carry meaning.

Alt text: Horizontal point estimates with confidence intervals compare three overall states for DHS/MICS and DHIS2. Fully aligned estimates are close, about 39% and 41%. DHS/MICS has a larger indeterminate share, about 44% versus 20%, while DHIS2 has a larger materially misaligned share, about 39% versus 17%.

Figure 2. D1-D4 domain-state distributions by data-source family. Design-weighted estimates and 95% confidence intervals are shown for aligned, materially misaligned and indeterminate or partially reported states in D1 data structure, D2 inferential goal, D3 method and D4 reporting or interpretation. Open and filled symbols distinguish DHS/MICS and DHIS2. Domains are displayed separately because they are linked within publications and are not treated as independent repeated observations.

Alt text: Four panels show A, M and I percentages for DHS/MICS and DHIS2. D2 is almost entirely aligned in both families. DHS/MICS has substantial insufficient reporting in D1 and D3. DHIS2 has more documented material misalignment in D3 and D4.

Figure 3. Prespecified Rwanda empirical illustration of survey-design consequences. Standard-error ratios for unweighted and weights-only analyses relative to the full survey design are shown for health-insurance prevalence and education-association coefficients in the 2019-20 and 2025 RDHS. The reference value of 1 corresponds to the analysis incorporating sampling weights, clusters and strata.

Alt text: Two panels show standard-error ratios for simplified analyses relative to the full survey design in the 2019-20 and 2025 Rwanda DHS. Open triangles denote unweighted analyses and open squares denote weights-only analyses. Values below 1 indicate narrower uncertainty under the simplified analysis. In 2025, unweighted and weights-only health-insurance prevalence

standard errors are about 62% and 68% of the full-design standard error; the corresponding higher-education odds-ratio shift is reported in the text rather than plotted.

References

1. Mbuagbaw L, Lawson DO, Puljak L, Allison DB, Thabane L. A tutorial on methodological studies: the what, when, how and why. BMC Med Res Methodol. 2020;20:226. doi:10.1186/s12874-020-01107-7.

2. Bell BA, Onwuegbuzie AJ, Ferron JM, Jiao QG, Hibbard ST, Kromrey JD. Use of design effects and sample weights in complex health survey data: a review of published articles using data from 3 commonly used adolescent health surveys. Am J Public Health. 2012;102:1399-1405. doi:10.2105/AJPH.2011.300398.

3. West BT, Sakshaug JW, Aurelien GAS. How big of a problem is analytic error in secondary analyses of survey data? PLoS One. 2016;11:e0158120. doi:10.1371/journal.pone.0158120.

4. Croft TN, Allen CK, Zachary BW, et al. Guide to DHS Statistics. Rockville: ICF; 2023.

5. Sheffel A, Wilson E, Munos M, Zeger S. Methods for analysis of complex survey data: an application using the Tanzanian 2015 Demographic and Health Survey and Service Provision Assessment. J Glob Health. 2019;9:020902. doi:10.7189/jogh.09.020902.

6. World Health Organization. Toolkit for analysis and use of routine health facility data: general principles. Geneva: World Health Organization; 2023.

7. World Health Organization. Implementation guide to the routine health information system toolkit. Geneva: World Health Organization; 2024.

8. Maiga A, Jiwani SS, Mutua MK, et al. Generating statistics from health facility data: the state of routine health information systems in Eastern and Southern Africa. BMJ Glob Health. 2019;4:e001849. doi:10.1136/bmjgh-2019-001849.

9. Farnham A, Loss G, Lyatuu I, Cossa H, Kulinkina AV, Winkler MS. A roadmap for using DHIS2 data to track progress in key health indicators in the Global South: experience from sub-Saharan Africa. BMC Public Health. 2023;23:1030. doi:10.1186/s12889-023-15979-z.

10. Wagenaar BH, Sherr K, Fernandes Q, Wagenaar AC. Using routine health information systems for well-designed health evaluations in low- and middle-income countries. Health Policy Plan. 2016;31:129-135. doi:10.1093/heapol/czv029.

11. Feng S, Hategeka C, Grepin KA. Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. Popul Health Metr. 2021;19:44. doi:10.1186/s12963-021-00274-z.

12. Amouzou A, Faye C, Wyss K, Boerma T. Strengthening routine health information systems for analysis and data use: a tipping point. BMC Health Serv Res. 2021;21 Suppl 1:618. doi:10.1186/s12913-021-06648-1.

13. Tuyishime AP, Muhirwa PK. Study protocol and registration. Open Science Framework. 2026. https://osf.io/89mk7/. Accessed 21 Oct 2026.

14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.

15. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61:29-48. doi:10.1348/000711006X126600.

16. National Institute of Statistics of Rwanda (NISR), Ministry of Health (MOH), ICF. Rwanda Demographic and Health Survey 2019-20 Final Report. Kigali and Rockville: NISR and ICF; 2021.

17. National Institute of Statistics of Rwanda (NISR), Ministry of Health (MOH), ICF. Rwanda Demographic and Health Survey 2025: Final Report. Kigali and Rockville: NISR, MOH and ICF; 2026.

18. Tuyishime AP, Muhirwa PK. EAC statistical alignment reproducibility package: derived data, code and audit materials. GitHub. 2026. Version 5f43ebbb08f9a8f02c883338d0caab5e3430f2b6. https://github.com/PrinceAudre/eac-statistical-alignment-reproducible/tree/5f43ebbb08f9a8f02c883338d0caab5e3430f2b6. Accessed 21 Oct 2026.

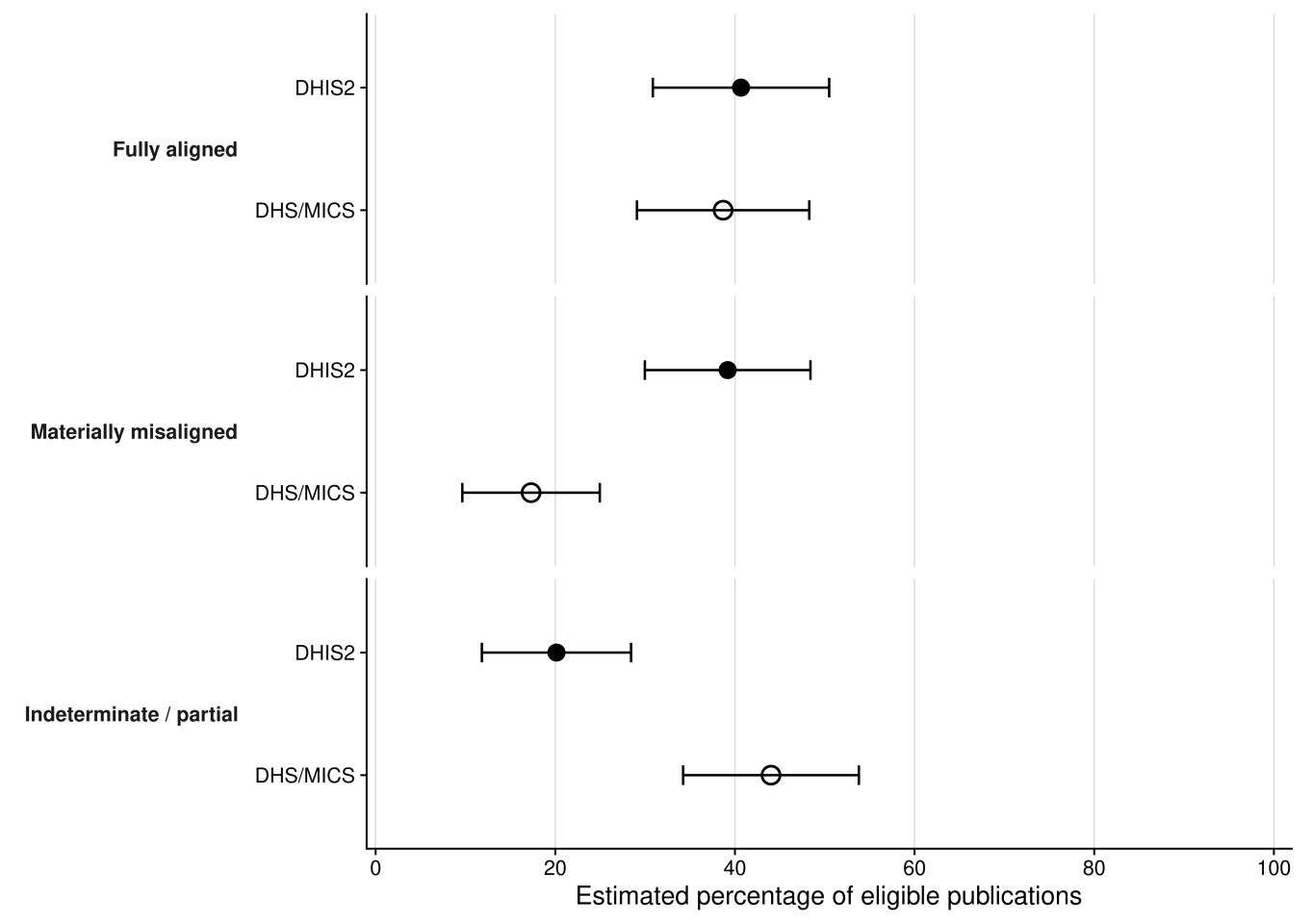
19. The DHS Program. Available datasets. https://dhsprogram.com/data/available-datasets.cfm. Accessed 21 Oct 2026.

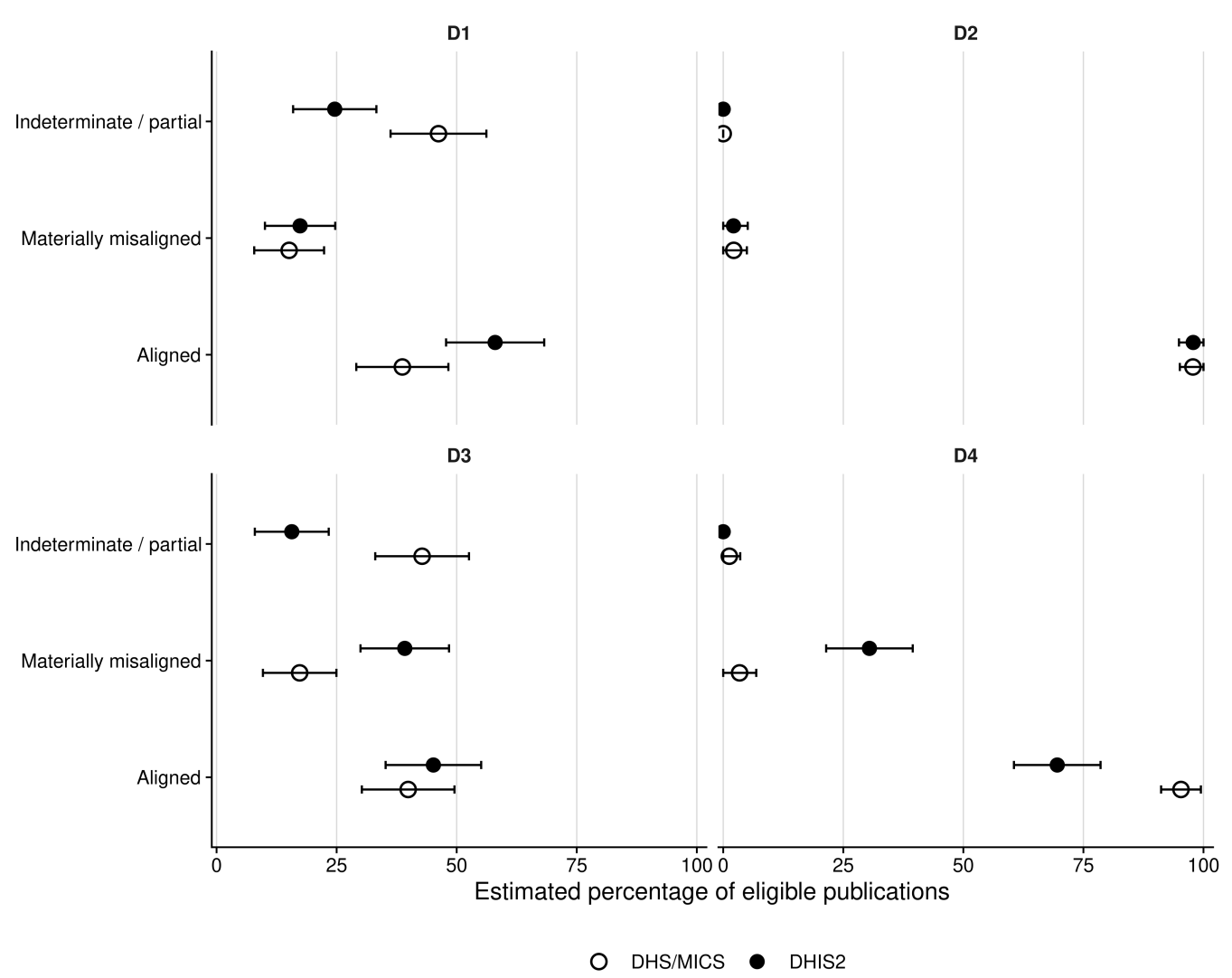
Additional files

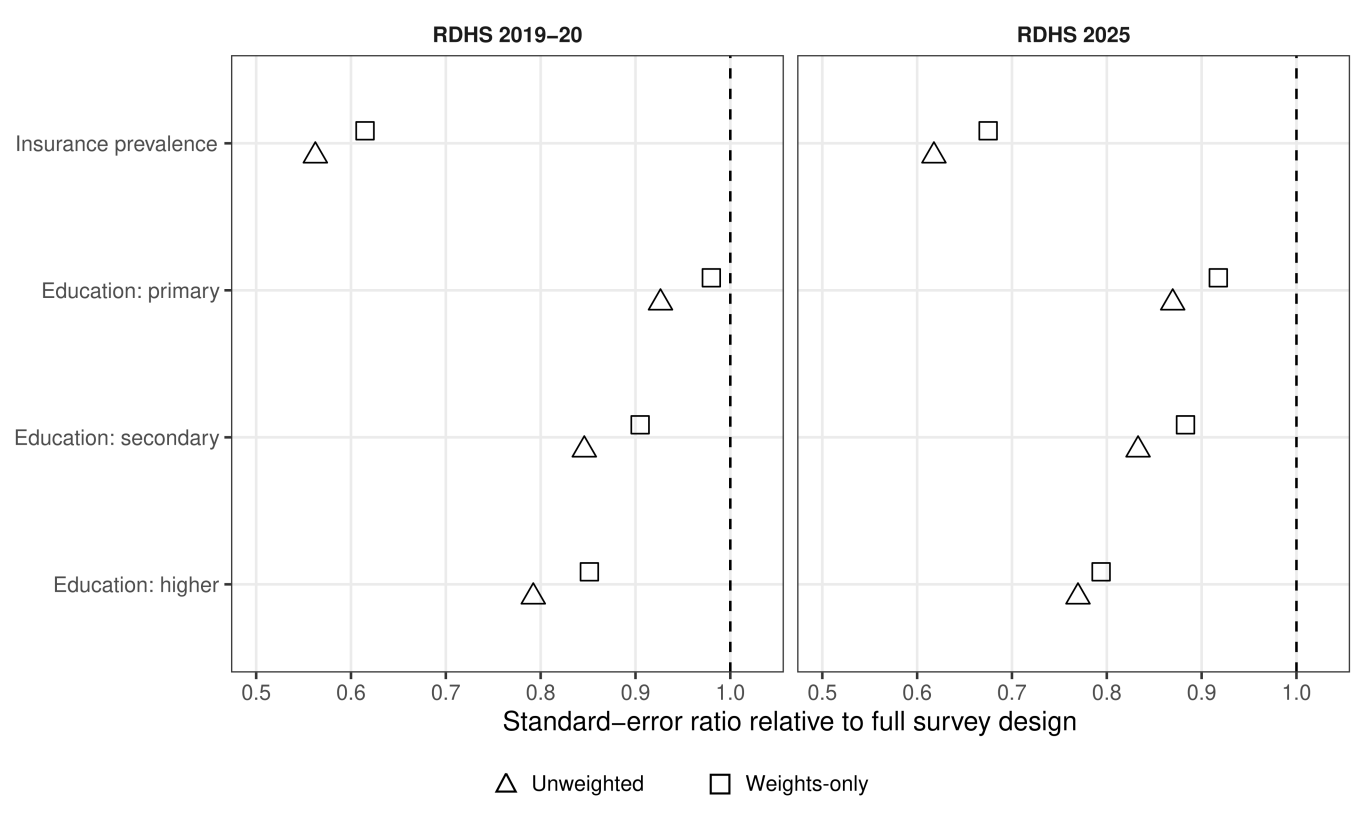
Additional file 1. Format: .xlsx. Title: Supplementary Tables S1-S8. Description: Family-by-country sampling strata and weights; full-text exclusion reasons; pre-adjudication reliability; sentinel indicators; sensitivity analyses; domain-state cross-tabulations; search-frame audit; and completed PRISMA 2020 checklist.

Additional file 2. Format: .pdf. Title: Supplementary Figure S2 - study flow. Description: Identification, screening, eligibility, detailed-coding probability sampling and independent reliability flow.

Additional file 3. Format: .pdf. Title: Supplementary Figure S1 - sentinel indicators. Description: Prespecified sentinel indicators by data-source family with 95% confidence intervals; applicable denominators and unclear states are reported in Supplementary Table S4 in Additional file 1.


Fully aligned
DHIS2
DHS/MICS
Materially misaligned
DHIS2
DHS/MICS
Indeterminate / partial
DHIS2
DHS/MICS
0
20
40
60
80
100
Estimated percentage of eligible publications


D1
D2
D3
D4
Indeterminate / partial
Materially misaligned
Aligned
0
25
50
75
100
Estimated percentage of eligible publications
DHS/MICS
DHIS2


RDHS 2019–20
RDHS 2025
Insurance prevalence
Education: primary
Education: secondary
Education: higher
0.5
0.6
0.7
0.8
0.9
1.0
Standard−error ratio relative to full survey design
Unweighted
Weights-only

Supplementary Files

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

- Additionalfile1.xlsx
- Additionalfile2.pdf
- Additionalfile3.pdf