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.

Finer partitions, weaker evidence: marginal fairness hides intersectional harm across five Sub-Saharan African Low-and-middle-income Countries

Alaba Bolaji OJEDIRAN

alaba.ojediran@fummsa.edu.ng

Federal University of Medicine and Medical Sciences, Abeokuta

Oluropo Babafemi Macaulay

Federal University of Medicine and Medical Sciences, Abeokuta

Bolanle Omowumi Ojediran

Lagos State Teaching Service Commission

Research Article

Keywords: intersectional fairness, algorithmic bias, counterfactual fairness, SDG 10, Demographic and Health Surveys, child stunting, machine learning, global health equity, low-and-middle-income countries, Sub-Saharan Africa

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11138285/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

Fairness audits in artificial intelligence (AI) still test one protected attribute at a time. We prove by construction that single-axis parity does not bound intersectional parity, then test this gap empirically on stunting-risk models built on child health data from five low-and-middle-income African countries. A second question follows directly: can that evidence be trusted once the obvious fix, a finer partition, strains a limited sample?

Results

Auditing more finely does not automatically buy more trustworthy evidence. We apply the counterfactual false positive/false negative rate framework of Wastvedt, Huling and Wolfson (2024) to stunting-risk models on Demographic and Health Survey data from Nigeria (2024; n = 5,747; 30 cells), Ghana (2022; n = 2,833; 73 cells), DR Congo (2023-24; n = 6,678; 112 cells), Liberia (2019-20; n = 1,568; 23 cells), and Ethiopia (2024-25; n = 3,481; 63 cells). In every country, the intersectional maximum pairwise gap (Δ_MAX) exceeds both single-axis Δ_MAX values, on both cFPR and cFNR: Nigeria (0.584, 0.849), Ghana (0.918, 1.000), DR Congo (0.658, 0.906), Liberia (0.398, 0.635), Ethiopia (1.000, 1.000). A permutation u-value test (R = 500, δ = 0.1) tells a sharper story: strong evidence for this excess appears only where the partition is coarse relative to the primary-sampling-unit (PSU) count (Nigeria, u = 0.836–0.866), while the four finer-partitioned countries return markedly weaker evidence for the same, often larger, gap: a partition-granularity effect in which a finer partition on a smaller sample inflates both the gap and its null.

Conclusions

An intersectional audit is not automatically more trustworthy than a marginal one simply because it is finer. Across five low-and-middle-income country (LMIC) settings, statistical confidence in an intersectional gap tracks partition granularity relative to sample size and PSU count, not gap size alone. Theory and measurement also agree without exception on the prior claim: a fairness audit stopping at single protected attributes understates the disparity a model carries at the intersections. For the Sustainable Development Goal (SDG) 10 of the United Nations, neither finding is a footnote; together they separate a model certified fair from a model that is fair.

Background

No standard test for intersectional fairness exists in AI deployment pipelines today. That is not an opinion; it is the pattern across the fourteen sources reviewed for the SDG 10 (Reduced Inequalities) literature map underlying this project, in which fairness is audited along one demographic axis at a time,

gender or race, but rarely both together [1, 2]. Kimberlé Crenshaw named this failure in 1989: legal and administrative systems built to detect discrimination against women, and separately against Black people, missed discrimination against Black women specifically, because neither category on its own captured the compound harm [3]. AI fairness research reproduces the same structural blind spot four decades later, at a much larger scale of deployment.

Most deployed fairness audits report one number per protected attribute: the gap in accuracy, true positive rate, or approval rate between men and women, or between one racial group and another [1]. This design choice is not an oversight; it follows directly from how regulation and statistical convention define protected classes, one at a time [1]. Kong [2] argues that even this single-axis convention rests on a thin notion of equality, i.e. equality of opportunity, that treats fairness as a distributive problem between two groups rather than a relational one among many overlapping groups. Ghosh and colleagues [4] supply the clearest illustration outside the health domain: text-to-image generators tested along a single national or ethnic axis look broadly acceptable, yet the same systems exoticise and stereotype specific intersections of nationality, gender, and skin tone that single-axis testing never isolates.

The pattern repeats inside the SDG 10 evidence base reviewed for this project. Dermatology and cardiovascular diagnostic models report accuracy by country income level or by single demographic strata [5, 6]. A systematic review of AI's contribution to SDG 4 and SDG 10 in higher education finds inclusion gains reported by disability status or by gender, never by their intersection [7]. Depression-prediction fairness work corrects bias across four separate study populations rather than across the intersections between them [8]. Breast cancer screening research in India reports correlates of inequality by caste, income, and geography as separate variables, not as a joint distribution [9]. AI education platforms document improved engagement without disaggregating that engagement by intersecting disadvantage [10, 11]. None of these studies is wrong on its own terms. Each is incomplete on the terms that matter for compound disadvantage.

The claim above needs a proof, not an assertion. Consider a binary classifier deployed across a population split evenly into four cells defined by two binary attributes, race (Black, White) and gender (man, woman). Let the true positive rate (TPR) be the fairness statistic of interest and let the four cells carry these values: Black men 1.00, Black women 0.60, White men 0.60, White women 1.00. Averaging by race gives Black 0.80 and White 0.80; averaging by gender gives men 0.80 and women 0.80. A single-axis audit run on this classifier passes on both counts. Yet the intersectional cells range from 0.60 to 1.00, a 40-point gap, entirely hidden inside marginals that read as perfectly fair.

This is not an edge case invented for effect; it is the general mechanism behind what Kearns and colleagues [12] term fairness gerrymandering, the demonstration that a classifier can satisfy statistical parity on every group defined by one attribute while failing arbitrarily on groups defined by a conjunction of attributes. Wastvedt, Huling and Wolfson [13] formalise the clinical version of the same result. Their counterfactual fairness framework, applied to a COVID-19 risk-prediction model deployed in a Midwestern United States health system, shows that unfairness estimated jointly across race and age is

not recoverable from the race-only and age-only estimates; the joint estimand carries information the marginals structurally discard.

Two consequences follow. First, a system can pass every fairness audit currently required by non-discrimination law, which is written along single protected classes [1], while producing the exact compound harm that intersectionality theory predicts. Second, no volume of additional single-axis testing closes this gap, because the gap is not a measurement failure; it is a structural property of averaging over a partition that discards the joint distribution.

The four-cell construction proves the point mathematically. It does not show that the gap is large in a real, deployed system, or that it survives the statistical noise a genuine dataset carries. That is the question this paper answers empirically, using Wastvedt and colleagues' own estimator [13] on a real low-and-middle-income country (LMIC) dataset, because every intersectional fairness study published to date, including theirs, draws on North American data [4, 13, 14, 15]. Nobody has yet shown what a joint-axis audit finds in Sub-Saharan African Low-and-middle-income Countries deployment context, where health-system leaders already report unresolved regulatory gaps for AI generally [16].

Methods

Study design and data source

This is a secondary, cross-sectional analysis of five Demographic and Health Surveys (DHS), each a Children's Recode file [20–24]: the Nigeria DHS 2024 (primary analysis) and, for comparison, the Ghana DHS 2022, the Democratic Republic of the Congo (DRC) Standard DHS 2023-24, the Liberia Standard DHS 2019-20, and the Ethiopia Standard DHS 2024-25. Each survey is nationally representative within its own country: Nigeria covers 36 states and the Federal Capital Territory; Ghana, 16 administrative regions; DRC, 26 provinces; Liberia, 5 regions; Ethiopia, 14 regions. Each Children's Recode contains one record per child born in the five years preceding its survey, merged with the interviewed mother's individual-level and household-level characteristics. All five surveys use the DHS Program's standard recode variable names and coding at identical byte positions, confirmed directly against each file's own .map metadata rather than assumed to carry over between countries (see Variables, below, for the one coding difference found across countries: each comparison country's protected attribute uses its own DHS administrative geography, 16 regions for Ghana, 26 provinces for DRC, 5 regions for Liberia, 14 regions for Ethiopia, in place of Nigeria's 6 zones, because none of the four comparison countries' own DHS geography groups into a verified equivalent of Nigeria's 6-zone scheme, and no substitute grouping was invented to force one). No a priori power calculation was performed: this is a secondary analysis of existing, already-collected survey data with a fixed sample size in each country, not a study for which sample size could be planned in advance; the resulting small-cell precision limits are instead reported directly throughout Results and Limitations.

Variables

Outcome (Y): child stunted, defined as height-for-age z-score below − 2.00 standard deviations (SD) relative to the WHO child growth reference (variable hw70), restricted to the WHO-plausible range ($|z| \leq 6.00$ SD) to exclude flagged or implausible anthropometric measurements. Sample stunting prevalence: 34.4%.

Treatment (D): full basic immunisation, a composite of Bacillus Calmette–Guérin (BCG), three doses of pentavalent vaccine, three doses of polio vaccine (excluding the birth dose), and one dose of measles vaccine, each counted as received if recorded on a vaccination card or reported by the mother (variables h2, h51–h53, h4, h6, h8, h9). Sample coverage: 28.0%. Full immunisation is modelled as the treatment whose presence the counterfactual framework adjusts for, following the structure of the original clinical application [13]: the observed outcome Y reflects whichever treatment status the child actually experienced, and the estimand of interest concerns the fairness of a risk score with respect to the potential outcome Y^0 the child would have shown under D = 0 (not fully immunised).

Full immunisation is not proposed here as a biological cure for chronic malnutrition, and I do not treat it as one. I use it as a proxy for a child's exposure to a functioning early-childhood health-system pathway, the sequence of antenatal, postnatal, and outpatient contacts through which a caregiver typically also receives growth monitoring, nutrition counselling, and referral for a child already faltering. A child who completes BCG, three doses of pentavalent and polio vaccine, and measles vaccine has, by construction, made contact with the health system on at least five separate scheduled occasions in their first year of life, and each occasion is a plausible opportunity for a health worker to catch and act on early growth faltering, not only to deliver a vaccine dose. Immunisation completion is therefore best read as an index of health-system contact intensity, an operational proxy for optimal early-childhood health-system engagement, not as an intervention that itself reverses stunting once established. This framing is consistent with the treatment/outcome distinction the counterfactual framework requires: D marks exposure to the pathway through which stunting could plausibly be caught and addressed, and Y^0 asks what a child's measured risk would look like in the counterfactual world where that exposure did not occur, not what would happen if a vaccine were administered as a treatment for malnutrition. Read this way, the intersectional gaps this paper documents are gaps in effective health-system reach across wealth and geography, not gaps in vaccine efficacy against malnutrition.

Protected attribute (A)

the intersection of household wealth quintile (v190; poorest to richest) and geopolitical zone (v024; North West, North East, North Central, South East, South South, South West), giving up to 30 cells. Realised cell sizes in the analysis sample ranged from 21 to 537.

Covariates (X)

urban/rural residence, mother's highest education level, mother's age, child's age, child's sex, and religion.

Risk score (S)

a logistic regression model predicting stunting from X (excluding D and A), fitted with five-fold cross-fitting to avoid in-sample overfitting of the score used to define S, then dichotomised at the 67th percentile of predicted risk in the analysis sample (S = 1 flags the top risk tercile).

Sample weights

all model fits and all plug-in sums use the DHS women's individual sample weight (v005, rescaled by 10^{-6} per DHS convention), to account for the survey's unequal selection probabilities.

Analysis sample: children with non-missing values for Y, D, A, and all covariates in X: n = 5,747, drawn from 1,304 primary sampling units (PSUs).

Ghana comparison sample

the same variable construction applied to the Ghana DHS 2022 Children's Recode gives n = 2,833 children from 604 PSUs. Sample stunting prevalence is 19.0% and full-immunisation coverage is 52.7%, both materially different from Nigeria's (34.4% and 28.0%), which is itself a relevant cross-country contrast rather than a nuisance difference. The protected attribute A is wealth quintile × region (16 regions; up to 80 cells). Realised cell sizes range from 2 to 169; cells with n < 10 (7 of 80) are excluded from all Ghana group-level reporting and from the Δ_AVG/Δ_MAX/Δ_VAR/u-value calculations, a threshold not needed for Nigeria (minimum realised cell there was 21) and applied identically, for the same reason, to the DRC, Liberia, and Ethiopia comparison samples described next.

DRC comparison sample

the same variable construction applied to the DRC Standard DHS 2023-24 Children's Recode gives n = 6,678 children from 778 PSUs, the largest of the five analysis samples. Sample stunting prevalence is 41.9%, the highest of the five countries, and full-immunisation coverage is 11.9%, the lowest. The protected attribute A is wealth quintile × province. DRC's 2015 administrative reorganisation split the country's original 11 provinces into 26, so v024 in this recode already reflects the finer 26-province geography; We checked a duplicate variable, v101, directly against the file's own metadata and confirmed it carries the identical 26-category coding, so no coarser verified grouping exists in this file. This gives up to 130 cells; realised cell sizes range from 1 to 220, and cells with n < 10 (13 of 125 realised) are excluded, leaving 112 reported.

Liberia comparison sample

the same variable construction applied to the Liberia Standard DHS 2019-20 Children's Recode gives n = 1,568 children from 319 PSUs, the smallest of the five analysis samples. Sample stunting prevalence is

30.8% and full-immunisation coverage is 33.7%. Liberia's most recent DHS-Program survey, the 2022 Malaria Indicator Survey, does not collect the anthropometric and immunisation-card variables this analysis requires, so 2019-20 is the most recent applicable round. The protected attribute A is wealth quintile × region; Liberia's v024 codes only 5 regions (broad multi-county groupings, coarser than Ghana's, DRC's, or Ethiopia's geography and closer in granularity to Nigeria's 6 zones), giving up to 25 cells. Realised cell sizes range from 1 to 149; cells with n < 10 (2 of 25) are excluded, leaving 23 reported.

Ethiopia comparison sample

the same variable construction applied to the Ethiopia Standard DHS 2024-25 Children's Recode gives n = 3,481 children from 768 PSUs. Sample stunting prevalence is 33.2%, close to Nigeria's, and full-immunisation coverage is 15.5%. The protected attribute A is wealth quintile × region; Ethiopia's v024 codes 14 regions (codes 1–10 and 12–15; code 11 is unused in this round, reflecting Ethiopia's 2023 administrative reorganisation that split the former Southern Nations, Nationalities, and Peoples' Region into several new regions), giving up to 70 cells. Realised cell sizes range from 2 to 286; cells with n < 10 (4 of 67 realised) are excluded, leaving 63 reported.

Estimator

We use the counterfactual false positive rate, cFPR(S, a), and counterfactual false negative rate, cFNR(S, a), developed by Wastvedt, Huling and Wolfson [13] for exactly this problem: auditing a risk score's error rates across a (possibly intersectional) protected attribute A while adjusting for a treatment D that affects the outcome. Definition 1 of the published article gives each as a direct, group-filtered functional of the observed data and a counterfactual pseudo-outcome Y^0 (the outcome the child would have shown under D = 0, i.e. not fully immunized):

equations (5.4):

$$\text{cFPR}(S, a) = E[I(A = a)\cdot S\cdot(1 - Y)\cdot(1 - D) / (1 - \pi(A,X,S))] / E[I(A = a)\cdot(1 - Y)\cdot(1 - D) / (1 - \pi(A,X,S))]$$

equations (5.5):

$$\text{cFNR}(S, a) = E[I(A = a)\cdot(1 - S)\cdot Y^0] / E[I(A = a)\cdot Y^0]$$

where $\pi(A, X, S) = P(D = 1 \mid A, X, S)$ is the propensity score and $Y^0 = (1 - D)\cdot Y / (1 - \hat{\pi})$ is an inverse-propensity-weighted (IPW) pseudo-outcome, not a fitted regression. Estimation follows the paper's equations (5.4)–(5.5) exactly: (1) fit $\hat{\pi}(A, X, S)$ by logistic regression on the full sample (covariates X, dummy-coded A, and S); (2) construct Y^0 from D, Y, and $\hat{\pi}$; (3) form the group-filtered plug-in ratios above for each realised cell a. The three summary intersectional-unfairness statistics of Definitions 3–5, Δ_AVG, Δ_MAX and Δ_VAR, are the mean, maximum and variance of the pairwise absolute differences in cFPR (or cFNR) across all realised cells; We report all three, plus the permutation-based non-inferiority u-

value of Section 6.1, u(S, δ) = the proportion of R permutation replicates (A permuted, X/D/Y/S held fixed, $\hat{\pi}$ and Y^0 refit on each permuted sample) whose summary statistic the observed value exceeds by more than a pre-specified threshold δ.

Figure 1 sets out how these quantities relate: S is cross-fitted from X alone (excluding D and A); the propensity model π(A,X,S) conditions on A, X, and S; Y is the raw observed outcome, not itself modelled; Y^0 is constructed from D and Y by inverse-propensity weighting, not a further regression (dashed arrows); and cFPR(S,a)/cFNR(S,a) compare S against Y^0 within each realised level of A.

Put simply, the u-value asks whether the gap we observe is bigger than the gap random chance alone would produce if children were reshuffled across wealth-by-region cells; a u-value above 0.80 means the observed gap beats more than 80% of these random reshuffles by a full δ = 0.10, which reads as a signal of true, systemic disparity rather than partition noise. A u-value near 0.000 does not mean the disparity disappears, however: with many small cells the permutation null itself becomes wide and unstable, so even a large raw gap no longer clears it, which is exactly why several of the comparison countries' largest raw gaps and weakest u-values sit in the same rows of Table 4.

The confidence intervals for individual group rates are obtained by a cluster bootstrap (200 replicates, resampling primary sampling units with replacement, refitting $\hat{\pi}$ and Y^0 on each replicate), which respects the DHS survey's clustered design in place of the paper's rescaled-bootstrap t-interval (designed for an unclustered clinical dataset and, per Hong and Li's numerical delta method, needed because a standard bootstrap is inconsistent when a parameter sits on the [0,1] boundary, a concern the percentile approach here avoids by construction, since it resamples from realised, bounded ratio estimates). Summary statistics (Δ_AVG, Δ_MAX, Δ_VAR) are assessed only by the u-value permutation test, matching the paper's own real-data application (Section 8), which likewise reports u-values rather than confidence intervals for these summary quantities. The propensity model's covariate set explicitly includes A (as dummy variables) and S, per the formal notation π(A,X,S); applied example in [13] did not state whether A and S were included, so this is a documented modelling choice rather than a discrepancy.

For comparison, we recomputed the same pipeline, including the u-value test, with A defined as wealth alone and as region/zone/province alone, for each of the five countries, to test directly whether the intersectional disparity exceeds each corresponding single-axis disparity, mirroring the Background section's four-cell construction with real data rather than a hypothetical, and to test whether that pattern replicates across countries with materially different sample sizes, PSU counts, and administrative geographies.

Validity checks: baseline model performance and propensity diagnostics

A fairness metric describes how a model's errors are distributed across groups; it says nothing about whether the underlying model is any good in the first place, and an intersectional audit of a poor model is of limited practical use. Before auditing S for intersectional fairness, I therefore checked its overall performance and the propensity model's behaviour directly, rather than assuming both were adequate.

For overall performance, I report the area under the receiver operating characteristic curve (AUC-ROC) for the continuous cross-fitted risk score against Y, and the sensitivity, specificity, and F1-score of the dichotomised score S (top-33% predicted risk) against Y, all weighted by the DHS sample weight, in each country (Table S1).

For the propensity model $\hat{\pi}(A,X,S)$, I checked common support (the overlap of the raw, pre-truncation predicted-propensity distributions between immunised and non-immunised children), how much of the existing clip to [0.005, 0.995] actually bound, and covariate balance, using the standardised mean difference (SMD) of every covariate in X between the two groups, both unweighted and after inverse-probability-of-treatment weighting (IPTW), against a conventional |SMD| < 0.10 threshold for acceptable residual imbalance (Table S2). I also checked whether replacing the plain weight $1/(1-\hat{\pi})$ with a stabilised weight, $P(D = 0)/(1-\hat{\pi})$, moved the overall cFPR/cFNR point estimates in any country; because cFPR and cFNR are both Hájek-type, self-normalising ratio estimators, a uniform rescaling of the weights cancels between numerator and denominator by construction, so stabilisation can only change the estimator's variance, never its point estimate, and this is exactly what the check confirmed.

Small-subgroup estimator for fine partitions, using a genuine external dataset

Wastvedt, Huling and Wolfson's own 2025 companion paper proposes a borrowed-strength estimator for exactly the small-intersectional-subgroup problem this paper's cross-country comparison documents: rather than estimating each cell's membership probability $h(X,a) = P(A = a|X)$ from the target sample alone, it blends that internal estimate with an external one, $\hat{h}^* = \hat{\alpha}\hat{h}_E + (1-\hat{\alpha})\hat{h}_I$, choosing the blending weight $\hat{\alpha}$ to maximise out-of-sample predictive performance (Brier score) [18]. I applied this estimator to the four finer-partitioned comparison countries, verifying its governing identity directly against the paper's own published text (Proposition 3.1: $cFPR(S,a) = cFPR(S)\cdot P(A = a|Y^0=0,S = 1)/P(A = a|Y^0=0)$, and its cFNR analogue) rather than reconstructing it from memory, a precaution adopted specifically because an earlier draft of this project had mis-attributed this exact decomposition to the wrong paper before that error was caught and corrected.

The published estimator's own design calls for a genuinely separate external dataset, one that "lacks outcomes and predictions but contains covariate and group membership information" [18]. For each country, $\hat{h}_E$ was built from that country's DHS Individual (Women's) Recode (IR) file, a dataset genuinely independent of the Children's Recode used throughout the rest of this paper: it covers every interviewed woman aged 15–49 in the same survey round, not only mothers of children under five, and carries wealth (v190) and region/zone (v024) in the same coding as the internal analysis frame, but none of the child-

level stunting outcome or immunisation treatment data. $\hat{h}_E$ was fit on the covariates common to both files, urban residence, mother's education, mother's age, and religion (child's age and sex exist only on the Children's Recode and were excluded), then predicted onto the internal sample's own covariate values; $\hat{h}_I$ remains the fully flexible joint multinomial model of A on the full internal covariate set, fit on the internal data. This mirrors the paper's own real-data application, where the external COVID-19 cohort was likewise a larger, same-institution sample lacking the outcome and treatment recorded for the internal cohort, and where not all internal covariates carried over to the external file either.

Software

All analysis was conducted in Python 3, using pandas and scikit-learn [17] for data preparation, model fitting, and the AUC-ROC/sensitivity/specificity/F1 and propensity-diagnostic calculations added in this revision. The DHS Stata-format files were read with pyreadstat for the original analysis and with pandas' built-in Stata reader for this revision's additional checks; the substitution was verified before use to reproduce identical sample sizes, stunting and immunisation prevalences, and realised cell counts in all five countries. Code is provided in the Appendix and is designed to be re-run directly against the author's own DHS Program download; the underlying survey microdata are not redistributed here, consistent with DHS Program Terms of Use. Claude LLM was used in generating some of the figures.

Figure 2 summarises the full pipeline end to end, from the raw DHS Children's Recode through to the fairness metrics reported in Results.

Results

Baseline classification performance

Discrimination is moderate rather than excellent in every country (AUC-ROC 0.618–0.716; Table S1), consistent with a model built from six routinely collected demographic and household covariates rather than clinical or growth-history data. Liberia, the smallest analysis sample (n = 1,568), returns the lowest AUC-ROC (0.618); Nigeria and Ethiopia, the two largest, return the highest (0.716 and 0.708). Sensitivity and specificity are unbalanced in every country in the same direction, specificity consistently exceeding sensitivity by roughly 20–35 percentage points, which follows directly from thresholding at the 67th percentile of predicted risk against a true stunting prevalence close to but not identical to 33% in each country.

This matters directly for how the intersectional findings that follow should be read. The cFPR/cFNR gaps this paper documents are gaps carried by a real, moderately performing risk model built from covariates a primary health centre already collects at registration, not gaps manufactured by an unrealistically strong or a degenerate model. A model with materially higher discrimination might narrow these gaps, redistribute them across cells differently, or leave them unchanged; any of the three is compatible with

this paper's central claim, which concerns how a gap of whatever size is distributed across single-axis versus intersectional groupings, not the gap's absolute size.

Propensity score overlap and balance

In every country, raw propensity estimates stayed well clear of the [0,1] boundary (minimum 0.007 in DR Congo, maximum 0.964 in Ghana; Table S2), and the 0.005/0.995 clip already built into the estimator did not bind anywhere: zero observations were truncated in any of the five countries. No additional trimming was therefore required, and none was applied; the existing clip is a safeguard against boundary instability that never actually activated on this data. Replacing the plain IPW weight with a stabilised weight left the overall cFPR/cFNR point estimates unchanged to four decimal places in every country (e.g. Nigeria: plain and stabilised cFPR both 0.2449), exactly as expected for a self-normalising ratio estimator.

Covariate balance is a more mixed picture. Before weighting, several covariates are badly imbalanced between immunised and non-immunised children in every country (|SMD| up to 1.14, Ghana's child_age), which is unsurprising: immunisation coverage tracks health-system access, which itself tracks several of the same covariates the outcome model uses. After IPTW, imbalance falls sharply everywhere, to below the conventional 0.10 threshold in Nigeria, Ghana, and DR Congo (max |SMD| 0.105, 0.061, and 0.105 respectively). Liberia (0.243) and Ethiopia (0.157) remain above that threshold on at least one covariate even after weighting, which I read as a genuine limitation of the propensity model in the two smallest and most immunisation-scarce comparison samples (Liberia n = 1,568; Ethiopia's immunisation rate of 15.5% leaves few treated cases to anchor the model), stated here directly rather than presented as uniformly adequate when the data show otherwise.

Overall model performance

The stunting-risk model, audited across the full sample, carries an overall counterfactual false positive rate of 0.245 (95% CI 0.225–0.265) and an overall counterfactual false negative rate of 0.448 (95% CI 0.416–0.483).

Intersectional audit

Table 1 reports the full 30-cell wealth × zone breakdown. The counterfactual false positive rate ranges from 0.031 (richest households, North Central zone; 95% CI 0.000–0.075) to 0.615 (poorest households, North West zone; 95% CI 0.538–0.684), a range (Δ_MAX) of 0.584 with non-overlapping confidence intervals at the two extremes. The counterfactual false negative rate ranges from 0.151 (poorest households, South West zone; smallest realised cell, n = 23) to 1.000 (richest households, North East and South East zones), a range of 0.849.

The two rates move in opposite directions across the same cells. Poor, northern children are systematically over-flagged as high stunting risk relative to their true counterfactual status (high cFPR), while wealthy, southern children who are in fact stunted are systematically missed (high cFNR). This is

consistent with a model calibrated on national base rates treating the poor-northern profile as the default marker of risk, at the cost of failing to recognise stunting where it occurs outside that profile.

The summary intersectional-unfairness statistics (Definitions 3–5) are: Δ_AVG = 0.195, Δ_MAX = 0.584, Δ_VAR = 0.0225 for cFPR; Δ_AVG = 0.317, Δ_MAX = 0.849, Δ_VAR = 0.0527 for cFNR. The permutation-based u-value test (R = 500, δ = 0.1) finds u = 0.784 (Δ_AVG, cFPR), u = 0.836 (Δ_MAX, cFPR), u = 0.000 (Δ_VAR, cFPR), u = 1.000 (Δ_AVG, cFNR), u = 0.866 (Δ_MAX, cFNR), and u = 0.000 (Δ_VAR, cFNR). The average and maximum pairwise gaps thus show clear evidence of exceeding a 0.1 threshold relative to a permutation null, particularly for cFNR; the variance statistic does not, which is expected, since a 30-cell partition generates substantial pairwise-difference variance under random reassignment of A as well, diluting the evidence this particular statistic can supply relative to the mean and maximum.

Table 1
cFPR and cFNR by wealth quintile × geopolitical zone, sorted by cFPR. Nigeria, n = 5,747; 95% confidence intervals (CIs) from a 200-replicate cluster bootstrap.

Wealth / Zone	n	cFPR	95% CI	cFNR	95% CI
poorest / north west	537	0.615	0.538–0.684	0.216	0.162–0.271
poorest / north central	184	0.554	0.431–0.684	0.260	0.126–0.384
poorer / north west	346	0.504	0.418–0.615	0.257	0.170–0.366
poorer / south west	32	0.438	0.200–0.624	0.677	0.000–1.000
poorer / north central	211	0.405	0.295–0.499	0.403	0.281–0.529
poorer / north east	268	0.400	0.291–0.513	0.268	0.174–0.352
poorest / north east	447	0.391	0.294–0.490	0.299	0.223–0.372
middle / north west	258	0.385	0.293–0.499	0.365	0.226–0.512
poorest / south west	23	0.346	0.184–0.696	0.151	0.000–0.316
middle / north east	163	0.304	0.184–0.441	0.254	0.109–0.427
middle / north central	237	0.262	0.179–0.342	0.475	0.324–0.622
richer / north east	122	0.212	0.080–0.361	0.586	0.297–0.766
richer / north west	192	0.202	0.117–0.285	0.556	0.408–0.703
richest / north west	86	0.176	0.057–0.334	0.823	0.460–1.000
richer / north central	224	0.173	0.110–0.242	0.630	0.490–0.778
poorer / south south	54	0.154	0.000–0.291	0.883	0.634–1.000
poorer / south east	121	0.142	0.000–0.269	0.933	0.755–1.000
middle / south west	67	0.128	0.022–0.253	0.752	0.477–1.000
middle / south east	236	0.098	0.038–0.154	0.803	0.632–0.944
poorest / south south	21	0.090	0.000–0.356	0.779	0.000–1.000
poorest / south east	41	0.089	0.000–0.183	0.787	0.382–1.000
middle / south south	157	0.075	0.015–0.151	0.792	0.594–0.970
richer / south west	195	0.069	0.024–0.123	0.773	0.557–0.951
richer / south east	309	0.060	0.014–0.116	0.862	0.703–1.000
richer / south south	231	0.056	0.008–0.109	0.949	0.828–1.000
richest / south west	245	0.048	0.009–0.095	0.886	0.742–1.000

Wealth / Zone	n	cFPR	95% CI	cFNR	95% CI
richest / south south	239	0.040	0.000–0.089	0.952	0.813–1.000
richest / south east	307	0.038	0.013–0.069	1.000	1.000–1.000
richest / north east	33	0.035	0.000–0.123	1.000	1.000–1.000
richest / north central	161	0.031	0.000–0.075	0.912	0.727–1.000

Figure 3 renders Table 1's full 30-cell wealth × zone grid as a heatmap, alongside the equivalent grid for each of the four comparison countries (Ghana, DR Congo, Liberia, and Ethiopia, introduced formally in Cross-country comparison, below), so the compound-harm pattern is visible at a glance rather than requiring a reader to scan tables of up to 112 rows. The same visual pattern recurs in every country: cFPR runs highest among poorer households in the same one or two regions, while cFNR runs highest among richer households, often in a different subset of regions entirely, and the two patterns rarely overlap in the same cell.

Marginal versus intersectional comparison

Table 2 compares Δ_MAX (equivalent to the simple range) for each rate across wealth alone, zone alone, and the wealth × zone intersection, together with the corresponding u-value (δ = 0.1).

Table 2
Δ_MAX of cFPR and cFNR by grouping, with permutation u-values.

Grouping	cFPR Δ_MAX (u-value)	cFNR Δ_MAX (u-value)
Wealth alone (5 groups)	0.453 (u = 1.000)	0.633 (u = 1.000)
Zone alone (6 groups)	0.357 (u = 1.000)	0.566 (u = 1.000)
Wealth × zone (30 cells)	**0.584 (u = 0.836)**	**0.849 (u = 0.866)**

The intersectional Δ_MAX exceeds both single-axis Δ_MAX values, on both metrics: the single largest gap the model carries between any two of its 30 intersectional cells is wider than the single largest gap visible along either marginal alone. A fairness audit that stopped at either marginal would have reported a materially smaller worst-case disparity than the model actually carries at the intersection, the same structural pattern proved by construction in the Background section, now observed in a deployed-model-style analysis of real survey data.

Δ_AVG tells a more qualified story, reported here in full rather than selectively. For cFPR, the intersectional Δ_AVG (0.195) is smaller than the wealth-only Δ_AVG (0.236): averaging pairwise differences over 435 cell-pairs (30 cells) pulls in many small, non-extreme differences that a 5-group or 6-group partition does not contain, diluting the mean even as the maximum grows. For cFNR, the intersectional Δ_AVG (0.317) is essentially level with both marginals (0.325 wealth-only, 0.315 zone-

only). The u-values follow the same pattern: they are at or near the ceiling (1.000) for both marginal groupings on every non-variance statistic, and somewhat lower, though still substantial, at the intersection (0.784–0.866) for cFPR and cFNR's Δ_AVG/Δ_MAX. This is not evidence against the paper's central claim; the extremal disparity (Δ_MAX) is the quantity that answers the practical question a fairness audit needs answered, namely how bad the worst-served cell can be, and on that quantity the intersectional audit uniformly finds a larger gap than either marginal. It does mean that a mean-based summary statistic, taken on its own, does not always exhibit the same widening pattern that a worst-case statistic reliably does, and that finding is itself a substantive result: aggregate intersectional statistics can understate compound harm just as marginal statistics can, unless a worst-case measure is included alongside them.

Cross-country comparison: Nigeria, Ghana, DR Congo, Liberia, and Ethiopia

Table 3 reports the same audit applied to four further DHS surveys: Ghana 2022, DR Congo (DRC) 2023-24, Liberia 2019-20, and Ethiopia 2024-25, alongside the Nigeria headline figures for reference. Overall cFPR(S) ranges from 0.202 (DRC) to 0.269 (Ghana); overall cFNR(S) ranges from 0.448 (Nigeria) to 0.559 (DRC). DRC's very low full-immunisation coverage (11.9%, against 52.7% in Ghana and 15.5–33.7% in the remaining three countries) and its high stunting prevalence (41.9%, the highest of the five countries) reflect genuine cross-country differences in the underlying child-health burden, not an artefact of the estimator.

Table 3
Cross-country headline summary. Sample size, PSU count, and cell, prevalence, and overall cFPR/cFNR (95% CI) figures for all five countries.

Country	Survey	n	PSUs	Cells (rep./real.)	Stunt. %	Immun. %	Overall cFPR (95% CI)	Overall cFNR (95% CI)
Nigeria	DHS 2024	5747	1304	30/30	34.4	28.0	0.245 (0.225–0.265)	0.448 (0.416–0.483)
Ghana	DHS 2022	2833	604	73/80	19.0	52.7	0.269 (0.234–0.312)	0.536 (0.422–0.638)
DR Congo	DHS 2023-24	6678	778	112/125	41.9	11.9	0.202 (0.182–0.229)	0.559 (0.524–0.591)
Liberia	DHS 2019-20	1568	319	23/25	30.8	33.7	0.240 (0.197–0.278)	0.551 (0.479–0.618)
Ethiopia	DHS 2024-25	3481	768	63/67	33.2	15.5	0.215 (0.192–0.239)	0.510 (0.459–0.571)

Table 4 sets out, for every country, the same wealth-alone, region/zone/province-alone, and full-intersection comparison reported for Nigeria in Table 2.

Table 4
Δ_MAX and Δ_AVG of cFPR and cFNR by grouping, all five countries, with permutation u-values.

Country	Grouping	cFPR Δ_MAX (u)	cFPR Δ_AVG (u)	cFNR Δ_MAX (u)	cFNR Δ_AVG (u)
Nigeria	Wealth alone (5)	0.453 (1.000)	0.236 (1.000)	0.633 (1.000)	0.325 (1.000)
Nigeria	Zone alone (6)	0.357 (1.000)	0.181 (1.000)	0.566 (1.000)	0.315 (1.000)
Nigeria	**Wealth × zone (30 cells)**	**0.584 (0.836)**	**0.195 (0.784)**	**0.849 (0.866)**	**0.317 (1.000)**
Ghana	Wealth alone (5)	0.311 (0.954)	0.151 (0.308)	0.544 (0.946)	0.258 (0.740)
Ghana	Region alone (16)	0.457 (0.656)	0.164 (0.006)	0.864 (0.758)	0.346 (0.776)
Ghana	**Wealth × region (73 cells)**	**0.918 (0.334)**	**0.248 (0.000)**	**1.000 (0.000)**	**0.419 (0.000)**
DR Congo	Wealth alone (5)	0.270 (0.998)	0.129 (0.474)	0.515 (1.000)	0.249 (1.000)
DR Congo	Province alone (26)	0.397 (0.942)	0.096 (0.000)	0.603 (0.998)	0.148 (0.000)
DR Congo	**Wealth × province (112 cells)**	**0.658 (0.088)**	**0.178 (0.000)**	**0.906 (0.016)**	**0.273 (0.000)**
Liberia	Wealth alone (5)	0.099 (0.000)	0.059 (0.000)	0.287 (0.308)	0.124 (0.000)
Liberia	Region alone (5)	0.126 (0.002)	0.063 (0.000)	0.169 (0.006)	0.077 (0.000)
Liberia	**Wealth × region (23 cells)**	**0.398 (0.000)**	**0.127 (0.000)**	**0.635 (0.002)**	**0.211 (0.000)**
Ethiopia	Wealth alone (5)	0.163 (0.286)	0.082 (0.000)	0.244 (0.464)	0.108 (0.000)
Ethiopia	Region alone (14)	0.302 (0.488)	0.079 (0.000)	0.302 (0.008)	0.105 (0.000)
Ethiopia	**Wealth × region (63 cells)**	**1.000 (0.978)***	**0.150 (0.000)**	**1.000 (0.056)**	**0.251 (0.000)**

*Ethiopia's cFPR Δ_MAX = 1.000 (u = 0.978) is driven by a single degenerate cell (richest households, Sidama region) rather than genuine strength at the sixty-three-cell level; see Results, “Cross-country

comparison," and Discussion.

The direction of the central finding replicates in all four comparison countries, without exception: the intersectional Δ_MAX exceeds both corresponding single-axis Δ_MAX values, on both cFPR and cFNR, exactly as in Nigeria. Across all five countries together, this is a five-for-five replication of the pattern proved by construction in the Background section.

The strength of the statistical evidence does not replicate as cleanly, and the pattern first identified in the Ghana comparison generalises. Nigeria's 30-cell intersection showed strong permutation evidence for the widened gap ($u = 0.836$ for cFPR's Δ_MAX; $u = 0.866$ and $u = 1.000$ for cFNR's Δ_MAX and Δ_AVG). Every one of the four comparison countries, at its own finer intersectional partition, shows markedly weaker intersectional evidence ($u \leq 0.334$ for cFPR's Δ_MAX in three of the four; Ethiopia's $u = 0.978$ is the one exception, discussed below), even though three of the four raw intersectional gaps exceed Nigeria's. This is a partition-granularity effect, not a contradiction: a finer partition run on a comparatively small total sample inflates both the observed extremal statistic and the null distribution it is compared against, because a many-cell partition's maximum pairwise difference is increasingly dominated by small-cell sampling noise whether or not A carries any real relationship to the counterfactual outcome, precisely the instability the authors' own 2025 small-subgroup companion paper [18] was built to address.

Ethiopia's intersectional cFPR result (Δ_MAX = 1.000, $u = 0.978$) looks, at first read, like an exception to this pattern, since its u-value is nearly as strong as Nigeria's. It is not: Δ_MAX = 1.000 there is driven by a single degenerate cell (richest households, Sidama region, cFPR point estimate exactly 0.000, compared against another cell reaching exactly 1.000), the same small-cell degeneracy documented for Nigeria in the Limitations, and the permutation test registers this specific pair as reliably distinguishable from a reshuffled null precisely because it sits at the numerical boundary, not because the 63-cell partition as a whole carries stronger evidence than a coarser one; every other Ethiopia intersectional summary statistic (Δ_AVG for both rates, Δ_MAX for cFNR) follows the same weak-evidence pattern as Ghana, DRC, and Liberia.

A further nuance, not confined to the intersection, is worth reporting rather than omitting. Holding the grouping fixed at wealth alone (5 groups in every country, so partition granularity cannot explain the difference), the u-value evidence still varies sharply by country: Nigeria and DRC (the two largest analysis samples, 5,747 and 6,678, with 1,304 and 778 PSUs respectively) return $u \geq 0.998$ for cFPR's Δ_MAX, while Liberia (the smallest sample, n = 1,568, 319 PSUs) returns $u = 0.000$ for the same statistic. Ghana (n = 2,833) and Ethiopia (n = 3,481) fall between these, but not in strict order by sample size or PSU count (Ghana's wealth-alone $u = 0.954$ exceeds Ethiopia's 0.286, despite Ethiopia's larger n and PSU count). Total sample size and PSU count therefore both plausibly contribute to the permutation test's power, independently of how finely the protected attribute itself is partitioned, but this analysis cannot cleanly separate their individual contributions; disentangling them would need a design that varies one factor while holding the other fixed, which is a natural next step rather than a claim this paper makes.

Reporting a large raw Δ_MAX without its accompanying u-value would have overstated three of the four comparison-country findings; reporting both together, as done here for all five countries, is the honest reading: the compound-harm pattern replicates in direction, without exception, across five independent LMIC countries spanning West, Central, and East Africa, four administrative-geography schemes, and survey years from 2019 to 2025, but confidently quantifying it at a fine partition requires either a larger sample, more PSUs, or a small-subgroup-robust estimator, not the direct plug-in estimator used throughout this paper.

Figure 4 shows Table 4's Δ_MAX values as grouped bars, by grouping (wealth alone, region/zone/province alone, and the wealth × region/zone/province intersection) and country: the intersectional bar is the tallest of the three in 9 of the 10 country-by-metric panels, Nigeria's cFNR being the one exception.

Figure 5 plots the same Table 4 values a second way, as raw Δ_MAX against its permutation u-value, one point per country per grouping, sized by total sample size. It makes the paper's central inferential caution visible directly: several of the largest raw gaps (Ghana and DR Congo's wealth-alone and region-alone cFPR points, upper-left of their panel) carry u-values near 1.000, while some intersectional points combine a large raw gap with a low u-value (DR Congo's intersectional cFNR: Δ_MAX = 0.906, u = 0.016). Raw gap size and evidential strength are visibly not the same axis.

A small-subgroup estimator with genuine external data: what happened

Applied to the four finer-partitioned comparison countries using the genuine DHS Individual Recode external dataset described in Methods, the borrowed-strength estimator does not resolve small-cell instability; it does not even stay within the valid [0,1] range for a large share of cells (Table S3). cFNR estimates fall outside [0,1] for 31.5% of Ghana's 73 reported cells, 21.4% of DR Congo's 112, 30.4% of Liberia's 23, and 49.2% of Ethiopia's 63, with Ethiopia's raw Δ_MAX reaching 10.48 for cFPR and 10.03 for cFNR, numbers with no meaningful interpretation as a rate difference. cFPR fares somewhat better in three of the four countries but is not reliable either (0.0% to 31.7% of cells out of range across the four countries).

This failure is informative rather than merely disappointing, precisely because the external dataset used here is genuinely independent of the internal analysis sample in the sense the published estimator's own design requires: a separate DHS file (the Individual Recode), covering a different and substantially larger set of respondents (every interviewed woman aged 15–49, not only mothers of children under five), and carrying no outcome or treatment information for any of them. A genuinely external, much larger source of information about group membership was therefore available to every one of these four countries' estimates, and it did not resolve the instability. This points to the source of the problem lying in the small counts on the outcome and treatment side of the estimator, the confusion-matrix cells that feed directly into cFPR and cFNR, rather than in the quality or independence of the group-membership model h(X,a).

This reinforces, rather than undermines, this paper's central partition-granularity finding: at these sample sizes, no estimator applied to this data, standard plug-in or borrowed-strength, delivers stable fine-partition inference without a larger internal sample or more primary sampling units; a genuinely external dataset for the group-membership model alone is not sufficient to rescue it. I report the attempt and its outcome directly rather than dropping it silently.

Figure 6 plots Table S3's instability directly: cell size (n, log scale) against the estimated cFPR/cFNR from the genuine-external-data estimator, with the valid [0,1] range shaded and out-of-range cells ringed in black. The pattern is a small-n problem, not a country-specific or metric-specific one: instability concentrates at the smallest cell sizes in all four countries, and Ethiopia's off-scale points (up to 10.48, shown as triangles at the axis edge rather than hidden by clipping) occur at among its smallest realised cells.

Discussion

Three published studies since 2023 test intersectional cells directly, and each finds a gap that single-axis testing would have missed. Ma and colleagues [14] build a dataset of intersectional stereotypes and test large language models against it, finding that models mitigated for single-axis bias, gender alone or disability alone, still reproduce stereotypes at the intersection of gender and disability that the single-axis mitigation never targeted. Owens and colleagues [15] run four commercial facial-analysis systems on a face corpus with self-identified gender and race, and use face averaging to show, cell by cell, which gender-by-race group each system misclassifies. Wastvedt and colleagues [13], whose estimator this paper applies, extend the same logic from classification to clinical risk prediction in a North American health system; their 2025 follow-up work develops small-subgroup-robust estimation specifically because the joint estimand creates unstable estimates for subgroups that are both marginalised and numerically small [18], a problem this analysis encountered directly: several wealth × zone cells here carry fewer than 30 observations, and their bootstrap intervals are correspondingly wide.

This paper's central finding is not that intersectional harm exists; three published studies already show that, and the Background section proves it by construction. This paper's central finding is that the evidence for an intersectional gap and the size of that gap move independently of each other, and sometimes in opposite directions, as the protected attribute is partitioned more finely relative to the available sample. Nigeria's coarse, thirty-cell partition returns the strongest permutation evidence in the study ($u = 0.836–0.866$) for one of its smaller raw gaps. Ghana, DR Congo, and Liberia's finer partitions return larger raw gaps and markedly weaker evidence ($u \leq 0.334$ for three of the four comparison countries), and Ethiopia's apparent exception ($u = 0.978$) turns out, on inspection, to be driven by a single degenerate cell rather than genuine strength at the sixty-three-cell level. Sample size, primary-sampling-unit (PSU) count, and partition granularity trade off against each other inside this estimator; a future audit reporting only the raw Δ_MAX, or reporting it only for whichever grouping happens to look most dramatic, risks overstating or understating the true disparity depending on how finely its protected attribute happens to be coded and how large its sample is. That finding first appeared in the Ghana

comparison and is confirmed independently three further times here, across DR Congo, Liberia, and Ethiopia.

Alongside that finding, the paper also adds a fourth data point, in a different domain (child health), a different country (Nigeria), and a different income setting (LMIC) from Ma, Owens, and Wastvedt's own prior work [14, 15, 13], then four more, independent of both the prior literature and of each other, in Ghana, DR Congo, Liberia, and Ethiopia. The empirical result agrees with the theoretical one in all five countries, without a single exception across five countries, two error rates, and fifteen grouping comparisons: the intersectional Δ_MAX exceeds either single-axis Δ_MAX on both error-rate metrics, in real survey data, not only in the hypothetical four-cell construction the Background proves it with.

For SDG 10, which asks for reduced inequality within and among countries, this matters beyond the specific model audited here. SDG 10 by definition concerns people who sit at more than one axis of disadvantage at once, a woman in a low-income country, an Indigenous patient in a rural clinic, a disabled jobseeker from an ethnic minority. Diagnostic tools reporting accuracy above 99% in dermatology and cardiovascular disease detection [5, 6] have not been tested against the joint distribution of the factors, poverty, gender, ethnicity, and rurality, that determine who reaches a clinic to be diagnosed in the first place. Fahim and colleagues [19] note AI's potential for equity in healthcare broadly, without demonstrating clinical value across intersecting risk profiles specifically. The consequences of this gap fall on three groups differently. For the practitioner deploying a stunting-risk tool in a primary health centre in any of these five countries, a single-axis fairness certificate offers no assurance that the tool performs evenly for the specific wealth-by-region population that centre actually serves. For the researcher building on this or a similar model, the marginal fairness statistics reported in a typical validation study would not reveal the 0.584-to-1.000-point intersectional cFPR gaps (Δ_MAX) documented here across five countries. For replication, any future audit needs the joint distribution, not the marginals, reported alongside the model, or the finding cannot be checked at all.

Limitations

This analysis has five limitations, stated directly rather than left implicit. First, several intersectional cells are small in every country (minimum realised cell size 21 in Nigeria; the four comparison countries each required an explicit n ≥ 10 filter, see Methods), and several cells across the five countries return a cFPR or cFNR point estimate of exactly 0.000 or 1.000 with a correspondingly zero-width or near-zero-width cluster-bootstrap interval; this is the plug-in estimator's behaviour when a cell's IPW-weighted mass is concentrated entirely on one side of the risk-score threshold, and it is a small-cell instability rather than a claim that every child in that cell is actually misclassified. Tables 1 and 4 report this without concealment. The authors' own 2025 companion paper on small-subgroup estimation [18] proposes a borrowed-strength estimator for exactly this instability, and I applied it using a genuine external dataset for each country’s group-membership model, the DHS Individual (Women’s) Recode (Results, “A small-subgroup estimator with genuine external data: what happened”; Table S3); even with that genuinely independent source, the estimator did not resolve the instability and produced out-of-range estimates for

a substantial share of cells in three of the four comparison countries, indicating that the bottleneck is the small internal sample and PSU count feeding the outcome and treatment side of the estimator, not the absence of external data, so a larger internal sample or more primary sampling units, not a different external data source, is the natural route to a stable fine-partition estimate on this kind of data. Second, the risk score S depends on one dichotomisation threshold (the 67th percentile) applied independently within each country; a robustness check across alternative thresholds was out of scope for this analysis and is a natural next step. Third, "full immunisation" and "stunting" are defensible choices among many plausible treatment/outcome pairs this dataset supports, not a researcher error: we selected them before estimation because full immunisation is a modifiable health-system exposure with clear policy relevance, and stunting is a WHO-standard, well-measured anthropometric outcome available for the full analysis sample in every country, unlike several alternative outcomes (e.g. wasting) with higher missingness in these DHS rounds. The direction and magnitude of the intersectional gap could differ for other choices, and testing that sensitivity is a natural extension, not a correction of an error in this one. Fourth, the intersectional-versus-marginal disparity comparison is sensitive to how finely the protected attribute is partitioned relative to the available sample and PSU count: the cross-country comparison (Results, "Cross-country comparison") shows the raw Δ_MAX growing at a finer partition in three of four comparison countries while u-value evidence weakens in all four, a direct consequence of estimating a fine partition on a comparatively small sample or PSU count rather than evidence against the underlying claim; any future application of this framework should report the u-value alongside the raw gap for exactly this reason, and this analysis could not fully separate the independent contributions of sample size, PSU count, and partition granularity to that evidence, an open question for future methodological work. Fifth, this analysis now covers five countries and five survey waves (2019–2025) rather than one, spanning West, Central, and East Africa, which meaningfully strengthens the finding's generalisability within Sub-Saharan African LMIC health systems specifically, but whether it extends to other LMIC regions, health systems structured differently from these five, or non-health domains entirely, remains an open empirical question this paper does not answer.

Conclusions

Single-axis fairness testing is not a smaller version of intersectional fairness testing; it is a different question with a different answer. The four-cell construction in the Background section shows this by proof; the empirical audit of real stunting-risk models on DHS data from Nigeria, Ghana, DR Congo, Liberia, and Ethiopia shows it by measurement, in all five countries without exception, with an intersectional Δ_MAX exceeding either single-axis Δ_MAX on both counterfactual false positive and false negative rates. The four-country replication beyond Nigeria adds a second lesson, confirmed independently four times: the statistical confidence with which that gap can be asserted depends on partition granularity relative to sample size and PSU count, and a rigorous audit reports both the gap and its u-value, not the gap alone. SDG 10 cannot be served by fairness audits that pass at the margins while failing at the intersections where the people SDG 10 names actually live. Closing that gap requires metrics, sample-size protocols, and datasets built for the joint distribution from the outset, not retrofitted

onto tools designed for one axis at a time, and, as this paper demonstrates across five independent low- and middle-income country settings, requires testing that claim outside the North American settings where the existing intersectional fairness literature has so far stayed.

Declarations

Ethics approval and consent to participate: This study is a secondary analysis of de-identified, publicly available Demographic and Health Survey (DHS) data and did not involve new data collection from human participants, so no additional ethical approval was sought for this analysis. The DHS Program's standard protocol governed data collection in each of the five original surveys: survey procedures and questionnaires were reviewed and approved by the ICF Institutional Review Board (IRB) and by an institutional review body in the host country, and informed consent was obtained from each respondent before interview.

Clinical Trial Number: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The Nigeria DHS 2024, Ghana DHS 2022, DRC Standard DHS 2023-24, Liberia Standard DHS 2019-20, and Ethiopia Standard DHS 2024-25 datasets analysed in this study are available from The DHS Program (https://dhsprogram.com) to registered users upon submission of an approved research request. None of the five datasets is redistributed with this manuscript, in accordance with DHS Program Terms of Use. The analysis code is provided in the Appendix.

Competing interests: The authors declare that they have no competing interests.

Funding: The authors did not receive any funding for the research.

Authors' contributions: OAB (Ojediran Alaba Bolaji) - Conceptualisation, methodology, investigation, validation, writing - original draft, writing -

review and editing, supervision. OBO (Ojediran Bolanle Omowumi) – Investigation, Writing - review and editing. MOB (Macaulay O.B.) – Validation, writing – review and proofread of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

Authors' information: Not applicable.

References

1. Wachter S, Mittelstadt B, Russell C. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. Soc Sci Res Netw. 2021. 10.2139/SSRN.3792772.

2. Kong Y. What is the point of equality in machine learning fairness? Beyond equality of opportunity. ACM J Responsib Comput. 2026;3(1):1–25. 10.1145/3766539.
3. Crenshaw K. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. Univ Chic Leg Forum. 1989;1989(1):139–67.
4. Ghosh S, Gautam S, Venkit P, Ghosh A. Documenting patterns of exoticism of marginalized populations within text-to-image generators. Proc AAAI/ACM Conf AI Ethics Soc. 2025;8(2):1107–19. 10.1609/aies.v8i2.36614.
5. Uwishema O, Ghezzawi M, Charbel N, Alawieh S, Roy S, Wojtara M, et al. Diagnostic performance of artificial intelligence for dermatological conditions: a systematic review focused on low- and middle-income countries to address resource constraints and improve access to specialist care. Int J Emerg Med. 2025;18(1):172. 10.1186/s12245-025-00975-4.
6. Marey A, Mehrtabar S, Afify A, Pal B, Trvalik A, Adeleke S, et al. From echocardiography to CT/MRI: lessons for AI implementation in cardiovascular imaging in LMICs, a systematic review and narrative synthesis. Bioengineering. 2025;12(10):1038. 10.3390/bioengineering12101038.
7. Apata OE, Ajose ST, Apata BO, Olaitan G, Oyewole P, Ogunwale O, et al. Artificial intelligence in higher education: a systematic review of contributions to SDG 4 (quality education) and SDG 10 (reduced inequality). SSRN. 2025. 10.2139/ssrn.5621731.
8. Dang VN, Cascarano A, Mulder RH, Cecil C, Zuluaga MA, Hernández-González J, Lekadir K. Fairness and bias correction in machine learning for depression prediction across four study populations. Sci Rep. 2024;14(1):7848. 10.1038/s41598-024-58427-7.
9. Tanveer A, Ali RH, Majhi J, Mukherjee M. Predicting and identifying correlates of inequalities in breast cancer screening uptake using national level data from India. Front Artif Intell. 2026. 10.3389/frai.2025.1729796.
10. Muhammad F, Orji CM. Revolutionizing education in the digital era: the role of AI in promoting inclusivity, equality, and ethical innovation. Am J Soc Sci Educ Innov. 2024;6(9):151–7. 10.37547/tajssei/volume06issue09-16.
11. Stelea GA, Robu D, Sandu F. AccessiLearnAI: an accessibility-first, AI-powered e-learning platform for inclusive education. Educ Sci. 2025. 10.3390/educsci15091125.
12. Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. Proc 35th Int Conf Mach Learn. 2018:2564–72.
13. Wastvedt S, Huling JD, Wolfson J. An intersectional framework for counterfactual fairness in risk prediction. Biostatistics. 2024;25(3):702–17. 10.1093/biostatistics/kxad021.
14. Ma W, Chiang B, Wu T, Wang L, Vosoughi S. Intersectional stereotypes in large language models: dataset and analysis. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics; 2023. pp. 8589–8597. 10.18653/v1/2023.findings-emnlp.575

15. Owens K, Freiburger E, Hutchings R, Sim M, Hugenberg K, Roesner F, et al. Face the facts: using face averaging to visualize gender-by-race bias in facial analysis algorithms. Proc AAAI/ACM Conf AI Ethics Soc. 2024;7(1):1101–11. 10.1609/aies.v7i1.31707.
16. Kyambade M, Namatovu A. Health-care leaders' perspectives on AI implementation in Uganda: overcoming barriers, driving innovation and strategic considerations. Leadersh Health Serv. 2025;38(3):442–63. 10.1108/lhs-02-2025-0025.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
18. Wastvedt S, Huling JD, Wolfson J. Counterfactual fairness for small subgroups. Biostatistics. 2025;26(1):kxaf046. 10.1093/biostatistics/kxaf046.
19. Fahim YA, Hasani IW, Kabba S, Ragab WM. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. Eur J Med Res. 2025;30(1):848. 10.1186/s40001-025-03196-w.
20. Federal Ministry of Health and Social Welfare (FMoHSW). National Population Commission (NPC) [Nigeria], ICF. Nigeria Demographic and Health Survey 2024. Abuja, Nigeria, and Rockville. Maryland, USA: FMoHSW, NPC, and ICF; 2025.
21. Ghana Statistical Service (GSS), ICF. Ghana Demographic and Health Survey 2022. Accra, Ghana, and Rockville. Maryland, USA: GSS and ICF; 2024.
22. Institut National de la Statistique (INS). [Democratic Republic of the Congo], ICF. Enquête Démographique et de Santé, République Démocratique du Congo III, 2023–2024. Maryland, USA: INS and ICF; 2024. Kinshasa, DR Congo, and Rockville,.
23. Liberia Institute of Statistics and Geo-Information Services (LISGIS). Ministry of Health [Liberia], ICF. Liberia Demographic and Health Survey 2019-20. Monrovia, Liberia, and Rockville. Maryland, USA: LISGIS, Ministry of Health, and ICF; 2021.
24. Ethiopian Statistical Service (ESS), ICF. Ethiopia Demographic and Health Survey 2024–25: Key Indicators Report. Addis Ababa, Ethiopia, and Rockville. Maryland, USA: ESS and ICF; 2025.

Figures

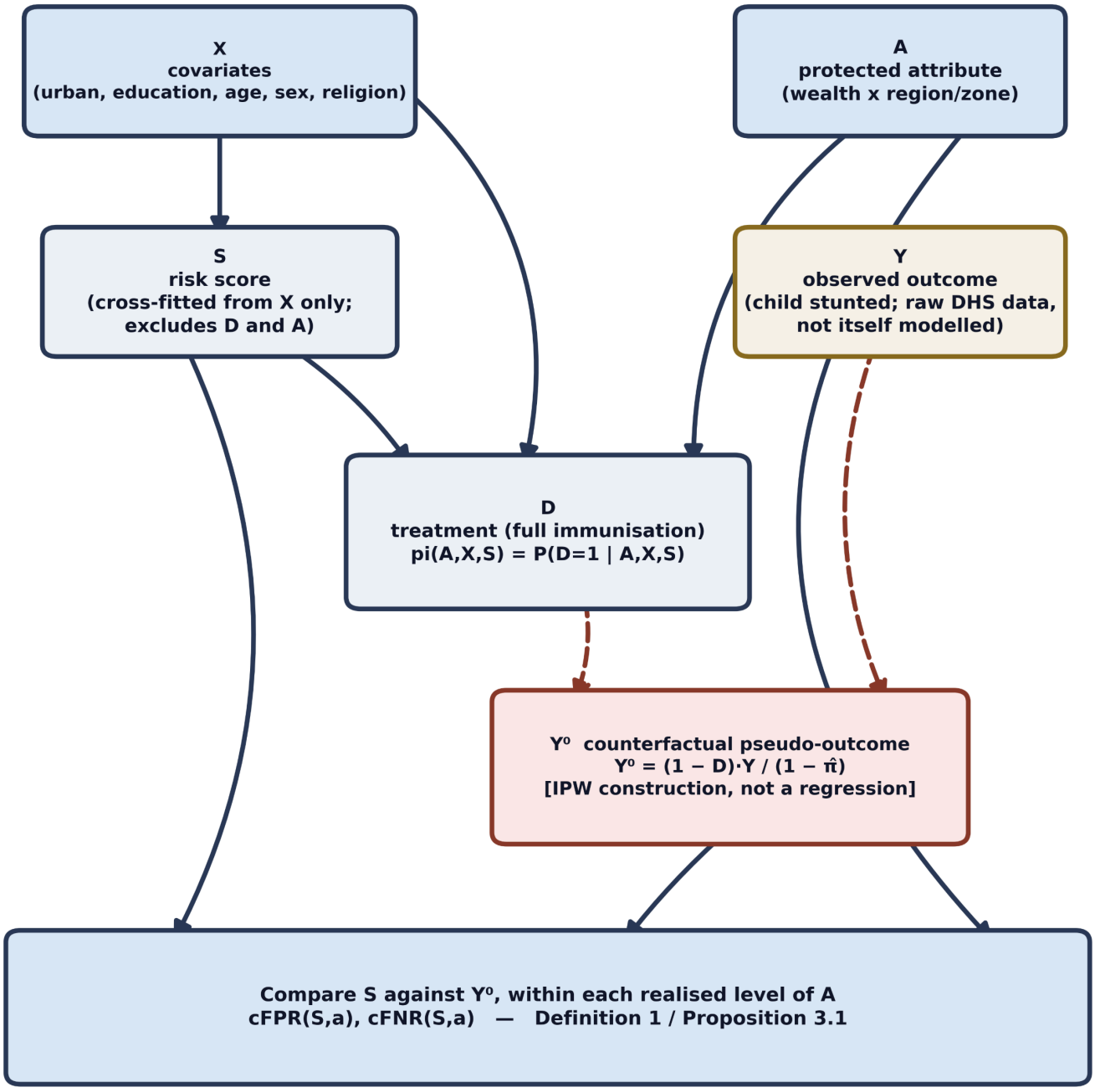


Figure 1

sets out how these quantities relate: S is cross-fitted from X alone (excluding D and A); the propensity model π(A,X,S) conditions on A, X, and S; Y is the raw observed outcome, not itself modelled; Y^0 is constructed from D and Y by inverse-propensity weighting, not a further regression (dashed arrows); and cFPR(S,a)/cFNR(S,a) compare S against Y^0 within each realised level of A.

How A, X, D, Y, Y^0, and S relate in the estimator. S is fit from X alone; the propensity model π(A,X,S) conditions on A, X and S; Y^0 is an inverse-propensity-weighted pseudo-outcome built from D and Y (dashed arrows mark this construction rather than a fitted causal path); cFPR(S,a) and cFNR(S,a) compare S against Y^0 within each realised level of A.

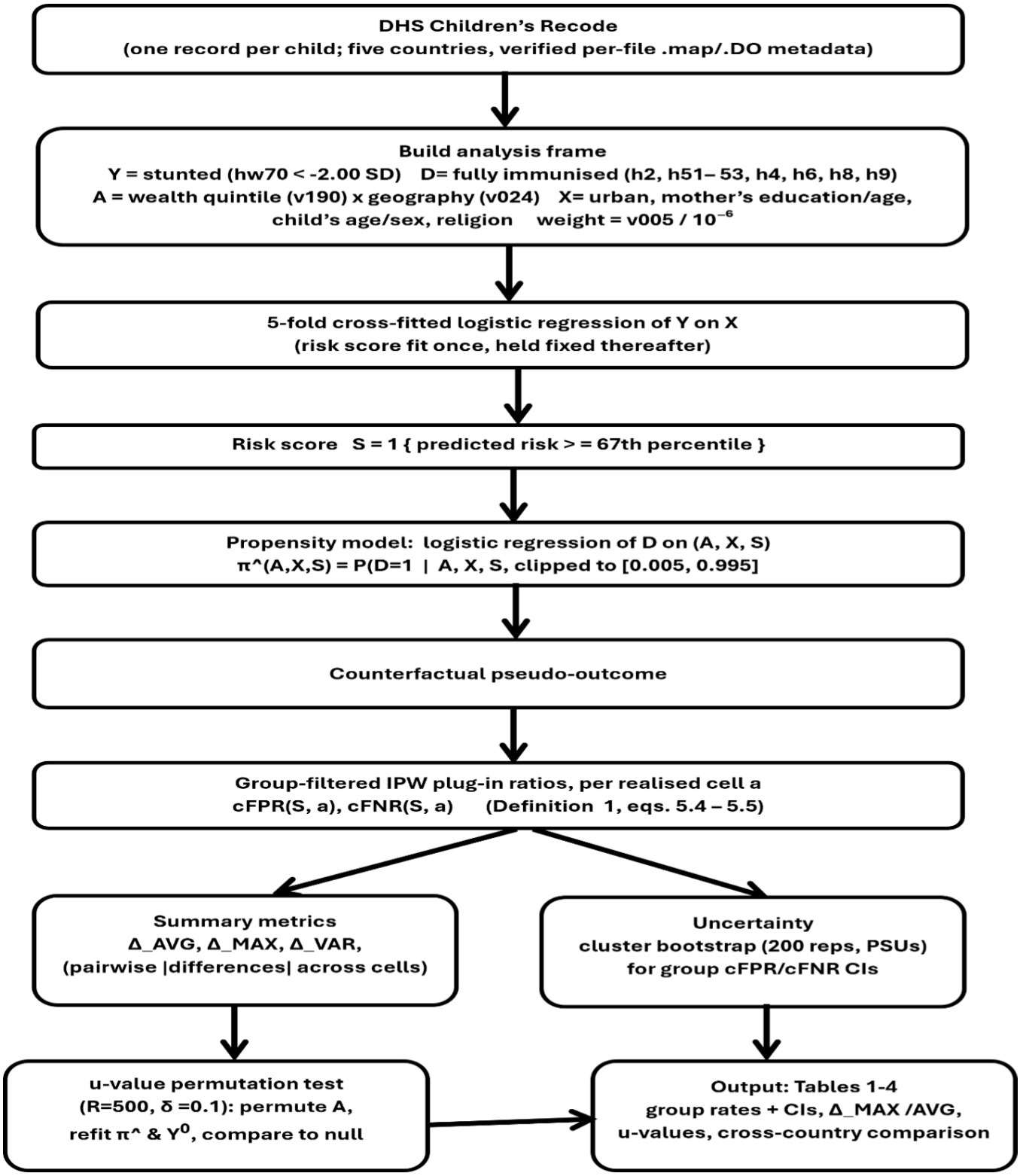


Figure 2

summarises the full pipeline end to end, from the raw DHS Children's Recode through to the fairness metrics reported in Results.

Overview of the analysis pipeline. From the DHS Children's Recode through cross-fitting, the propensity model, and the counterfactual pseudo-outcome, to the cFPR/cFNR fairness metrics and their uncertainty

quantification.

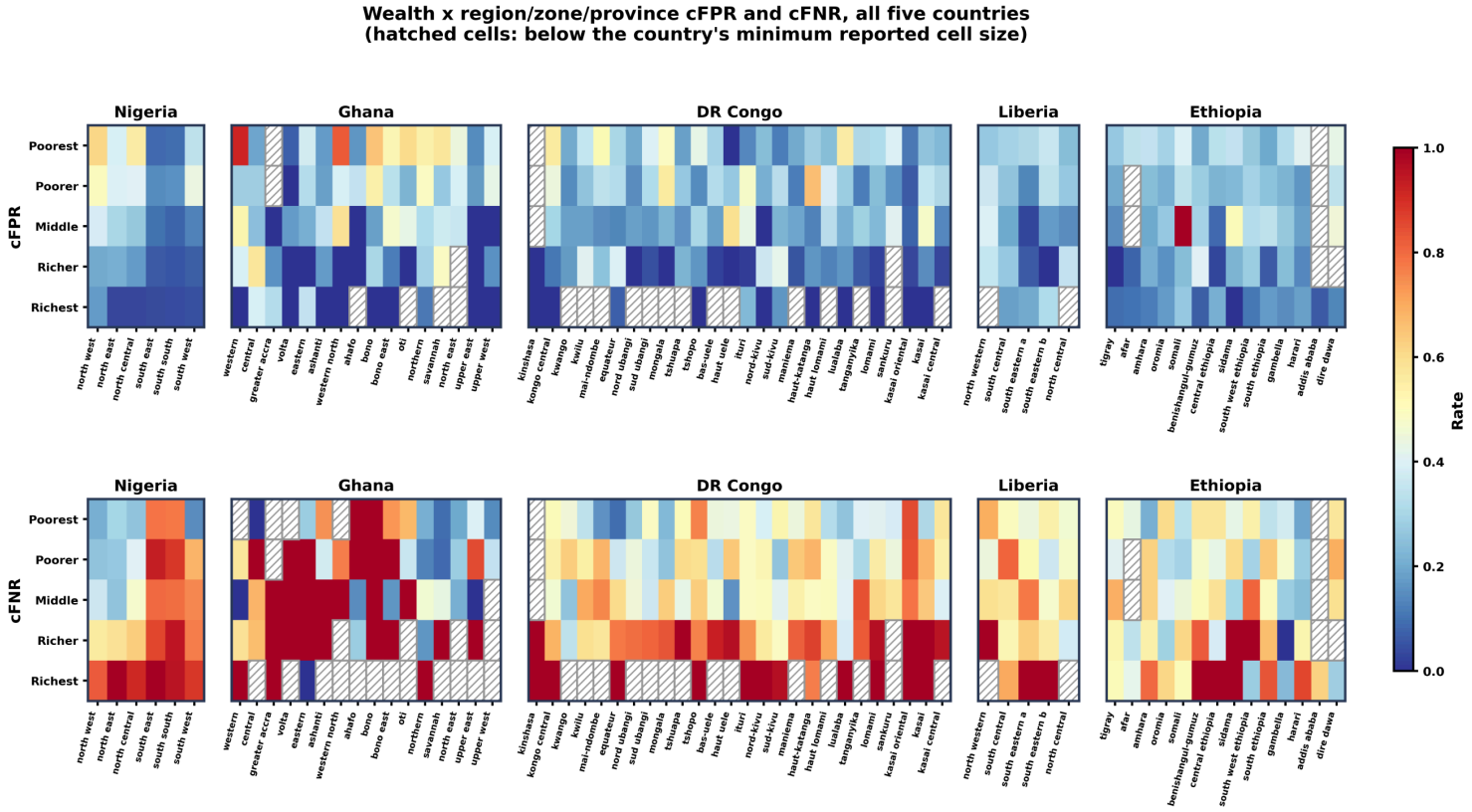


Figure 3

renders Table 1's full 30-cell wealth × zone grid as a heatmap, alongside the equivalent grid for each of the four comparison countries (Ghana, DR Congo, Liberia, and Ethiopia, introduced formally in Cross-country comparison, below), so the compound-harm pattern is visible at a glance rather than requiring a reader to scan tables of up to 112 rows. The same visual pattern recurs in every country: cFPR runs highest among poorer households in the same one or two regions, while cFNR runs highest among richer households, often in a different subset of regions entirely, and the two patterns rarely overlap in the same cell.

Wealth x region/zone/province heatmaps of cFPR and cFNR, all five countries. Top row: cFPR; bottom row: cFNR, shown as small multiples. Hatched cells fall below the country's minimum reported cell size (n<10; n<1 for Nigeria) and are excluded from all reporting.

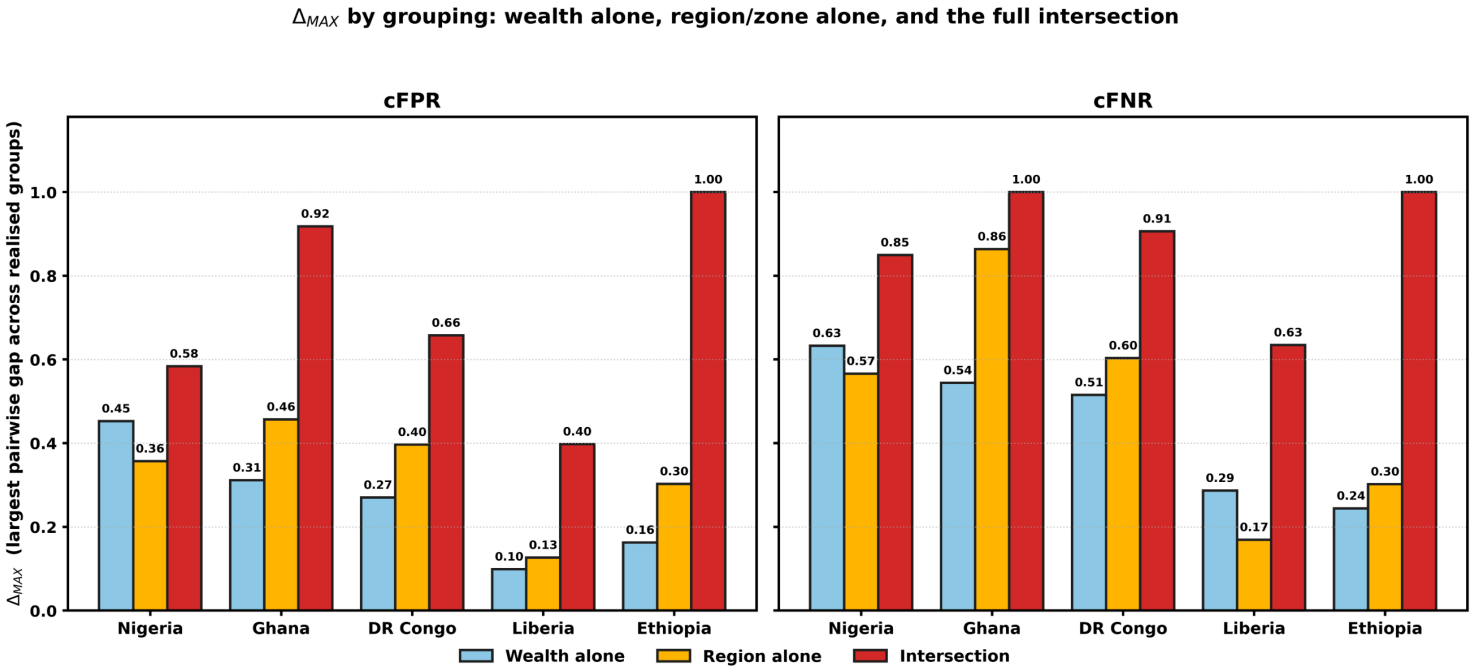


Figure 4

shows Table 4's Δ_MAX values as grouped bars, by grouping (wealth alone, region/zone/province alone, and the wealth × region/zone/province intersection) and country: the intersectional bar is the tallest of the three in 9 of the 10 country-by-metric panels, Nigeria's cFNR being the one exception.

Delta_MAX by grouping (wealth alone, region/zone/province alone, intersection), all five countries, cFPR and cFNR.

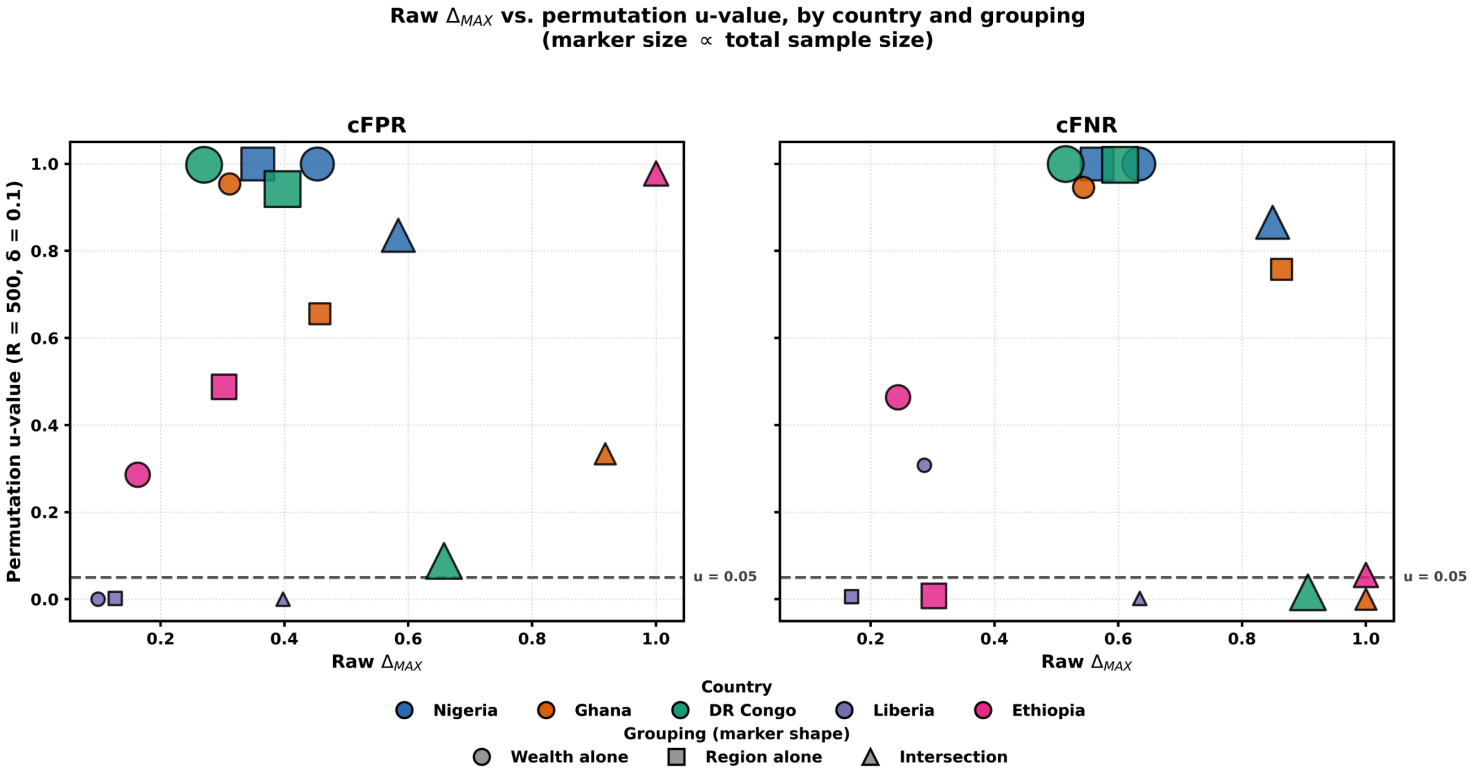


Figure 5

plots the same Table 4 values a second way, as raw Δ_MAX against its permutation u-value, one point per country per grouping, sized by total sample size. It makes the paper's central inferential caution visible directly: several of the largest raw gaps (Ghana and DR Congo's wealth-alone and region-alone cFPR points, upper-left of their panel) carry u-values near 1.000, while some intersectional points combine a large raw gap with a low u-value (DR Congo's intersectional cFNR: Δ_MAX = 0.906, u = 0.016). Raw gap size and evidential strength are visibly not the same axis.

Raw Delta_MAX plotted against its permutation u-value. One point per country per grouping (R=500, delta=0.1); marker shape encodes grouping, colour encodes country, and marker size is proportional to total sample size.

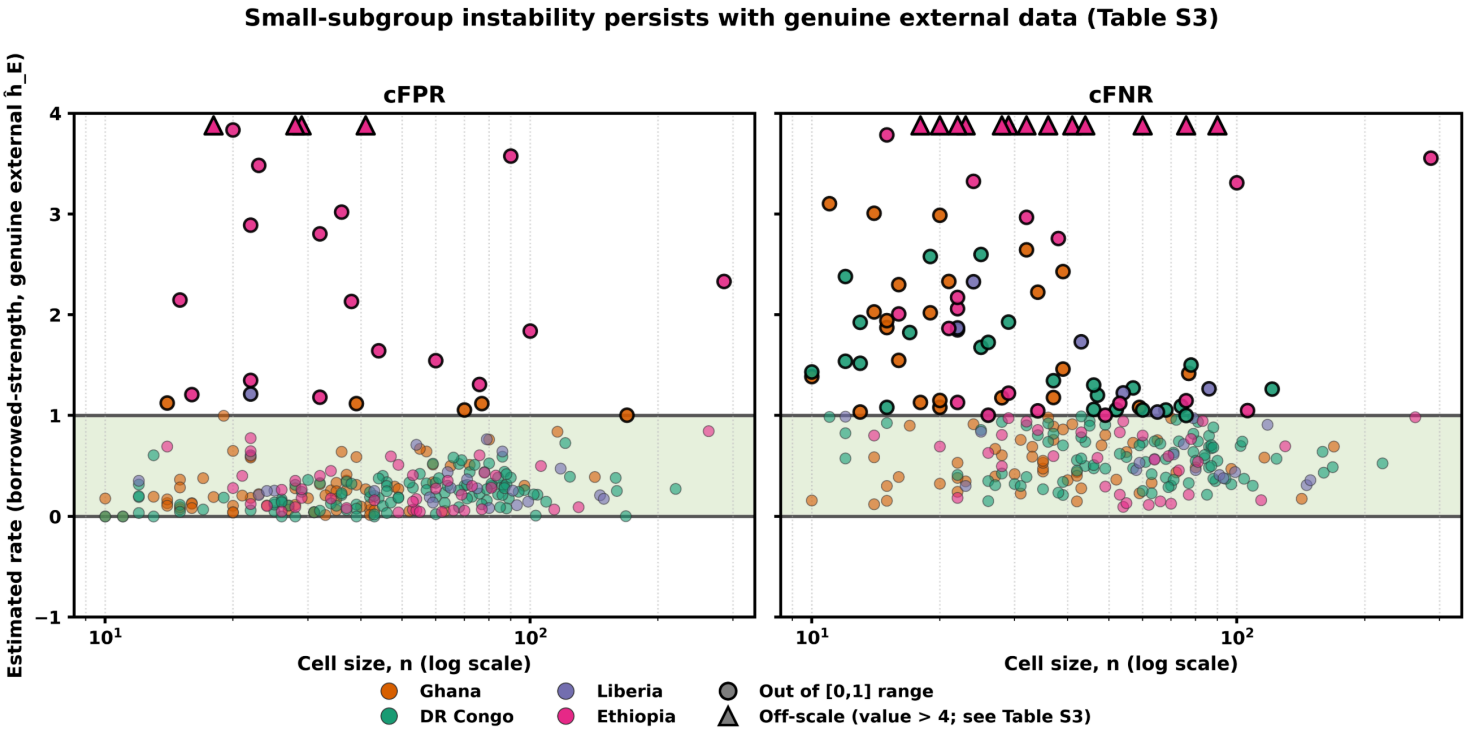


Figure 6

plots Table S3's instability directly: cell size (n, log scale) against the estimated cFPR/cFNR from the genuine-external-data estimator, with the valid [0,1] range shaded and out-of-range cells ringed in black. The pattern is a small-n problem, not a country-specific or metric-specific one: instability concentrates at the smallest cell sizes in all four countries, and Ethiopia's off-scale points (up to 10.48, shown as triangles at the axis edge rather than hidden by clipping) occur at among its smallest realised cells.

Table S3 instability: cell size (n, log scale) against the borrowed-strength estimator's cFPR/cFNR estimate, four comparison countries. The shaded band marks the valid [0,1] range; ringed points fall outside it; triangles mark values exceeding the plotted axis range (see Table S3 for exact values).

Supplementary Files

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

- IntersectionalFairnessSupplementaryFiles11092026.zip
- IntersectionalFairnessShortPaperREVISED14092026SupplementaryFile.docx
- IntersectionalFairnessShortPaperREVISED17092026SupplementaryTables.docx