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.

Interpreting NHANES CKD-Compatible Models under Scarcity

Hanif Abdulazeez

abdulazeezhanif00@gmail.com

Abubakar Tafawa Balewa University

Jamiu Ogirima Bello

Research and Development Department, Otabel Biomedical Solutions Ltd.

Abdulmumeen Adekunle Issa

Abubakar Tafawa Balewa University

Ridwanullah Eyitayo Abdulazeez

Abubakar Tafawa Balewa University

Abimbola Musa Yusuf

National Orthopaedic Hospital, Igbobi, Lagos

Kazeem Eyitayo Lasisi

Abubakar Tafawa Balewa University

Research Article

Keywords: CKD

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11335077/v1

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

Additional Declarations: No competing interests reported.

Interpreting NHANES CKD-Compatible Models under Scarcity

Hanif Abdulazeez[1] — abdulazeezhanif00@gmail.com — ORCID 0009-0005-3230-5022 — *Corresponding author*

Jamiu Ogirima Bello[2] — ORCID 0009-0002-6687-0044

Abdulmumeen Adekunle Issa[1] — ORCID 0000-0002-9075-1750

Ridwanullah Eyitayo Abdulazeez[3] — ORCID 0009-0006-4910-0485

Abimbola Musa Yusuf[4]

Kazeem Eyitayo Lasisi[1] — ORCID 0000-0001-9306-5994

Affiliations:

[1] Department of Statistics, Faculty of Science, Gubi Campus, Tafawa Balewa Way, P.M.B. 0248, Bauchi, Bauchi State, 740004, Nigeria.

[2] Research and Development Department, Otabel Biomedical Solutions Ltd, KM 2, Airport-Ningi Road, Bauchi, Bauchi State, Nigeria.

[3] Department of Biochemistry, Faculty of Science, Gubi Campus, Tafawa Balewa Way, P.M.B. 0248, Bauchi, Bauchi State, 740004, Nigeria.

[4] National Orthopaedic Hospital, Igbobi, Lagos (NOHIL), 120/124 Ikorodu Road, Igbobi, P.M.B. 2009, Yaba, Lagos State, 102216, Nigeria.

Abstract

Background: To evaluate whether four interpretability methods (standardized logistic regression coefficients, permutation importance, SHAP [Shapley Additive exPlanations], and LIME [Local Interpretable Model-agnostic Explanations]) produce consistent predictor rankings for a National Health and Nutrition Examination Survey (NHANES)-defined chronic kidney disease (CKD)-compatible abnormality as training sample size decreases, and whether predictive performance remains a reliable proxy for interpretability stability under data scarcity.

Methods: Using NHANES data (N=31,958 analytic pool), logistic regression models were fit on 100 repeated draws at each of seven sample sizes (N=5,000 to N=50). Model/draw validity and method-specific computational failure were assessed separately. Stability was measured via top-5 Jaccard similarity and full-ranking Spearman correlation, with formal pairwise significance testing (false discovery rate (FDR)-corrected) across all four methods. Held-out area under the receiver operating characteristic curve (AUC) and Brier score were evaluated via an 80/20 split.

Results: Interpretability stability declined sharply (mean Spearman 0.907→0.429, N=5,000→N=250) while predictive performance remained comparatively stable over the same range (AUC 0.776→0.748). Of 72 formal method-pair comparisons, 36 were statistically significant; standardized-β had the lowest Spearman stability at every rung, while SHAP was the most consistently distinguishable and, on point estimates, the most stable. Predictor-level vulnerability varied substantially: age remained stable (96.9–100.0% top-5 appearance) while BMI (21.0%→78.1%) and diabetes-borderline status (18.0%→56.3%) were unstable. BMI overshot the 71.4% chance baseline, and diabetes-borderline's rate differed sharply by method (standardized-β 60%→83%, SHAP 1%→33%).

Conclusion: Interpretability method rankings degrade substantially under data scarcity even when predictive performance does not, and this degradation is not uniform across methods or predictors. Predictive performance alone cannot indicate whether a small-sample model's interpretability output is reproducible; method choice, predictor prevalence, and sample size jointly determine whether it can be relied upon.

Keywords:

1. CKD
2. Interpretability
3. SHAP
4. LIME
5. Permutation importance
6. Sample size
7. Reproducibility

1. Background

1.1 Background of the Study

Kidney disease affects an estimated 674 million people worldwide as of 2021, with the burden falling disproportionately on low- and middle-income settings where limited health expenditure and nephrology access compound CKD's clinical and economic toll [1,2]. CKD research from these settings is itself resource-constrained: a systematic review of 90 sub-

Saharan African CKD studies found only 20% medium-quality and 3% high-quality by standard criteria, driven largely by small, non-population-based samples [3]. Statistical and machine-learning models are increasingly proposed for low-cost CKD risk identification in these settings, but their clinical value depends on whether identified risk factors are reproducible, not just predictive accuracy.

Interpretability methods split into two kinds: model-derived measures (standardized coefficients, permutation importance) and post-hoc local explanation methods (SHAP, LIME) that approximate model behavior and aggregate to a global ranking. Post-hoc methods often fail a basic robustness test, as near-identical inputs can yield substantially different explanations [4], but how any of these methods behave at only a few hundred or few dozen observations remains largely unvalidated.

1.2 Statement of the Problem

This scarcity problem is not addressed by existing interpretability-method comparisons, which test a single method in isolation, evaluate robustness only to small perturbations at a fixed sample size, or test sample-size ladders without a multi-method, explicit failure-rate comparison [4,5,6] (Related Work, Section 2). Whether these methods retain consistent, actionable predictor rankings as sample size falls into the hundreds or fewer, and whether outright ranking failure should itself be measured, has not been systematically examined.

1.3 Objectives of the Study

To address this gap, the present study set out to:

1. Compare the stability of four interpretability methods (standardized logistic regression coefficients, permutation importance, SHAP, and LIME) in identifying important

predictors of NHANES-defined CKD-compatible abnormality risk across a seven-rung sample-size ladder (N=5,000 to N=50).

2. Quantify model/draw validity failure rates across the ladder, and separately assess whether the four methods exhibit independent computational failure once a valid model exists.
3. Assess consistency of top-ranked predictors within each method using Jaccard similarity and Spearman rank correlation across repeated draws.
4. Identify which predictor categories are most vulnerable to instability as sample size decreases.
5. Evaluate held-out predictive performance (AUC, Brier score) across the same ladder, and its relationship to interpretability stability.

1.4 Scope of the Study

BMXWAIST (waist circumference) was excluded for high collinearity with BMXBMI (r = 0.9055), which inflates coefficient variance and destabilizes permutation-based importance (Section 2.2).

Statement of Significance

Problem or Issue	Interpretability methods (SHAP, LIME, permutation importance, standardized coefficients) are increasingly proposed for low-cost CKD risk prediction in resource-constrained settings, but their reliability when only small samples are available remains largely unvalidated.
What is Already Known	Existing comparisons test a single method in isolation or evaluate robustness only at one fixed sample size; to our knowledge, none combines a full sample-size ladder, a four-method comparison, and an explicit, reproducible failure-rate metric.

What This Paper Adds	A seven-rung sample-size ladder (N=5,000 to N=50) testing all four methods under identical conditions, formally separating model/draw validity failure from method-specific computational failure, testing between-method differences formally, and showing that ranking stability falls sharply while held-out discrimination changes little until N=100.
Who Would Benefit	Researchers and clinicians building or evaluating interpretability tools for clinical prediction models in small-sample health data settings

Related Work

Scope and Approach to the Review

This review covers three bodies of work bearing on the stability gap this project addresses: interpretability-stability testing in a clinical cohort, cross-method and CKD-specific SHAP/LIME/permutation-importance comparisons, and precedent for testing prediction-model stability across a sample-size range. Sources were included where they combine at least two of: a head-to-head comparison of multiple interpretability methods, sample size as an explicit experimental axis, or a reproducible, rule-based failure metric.

Prior Precedent for Stability Testing and Method Comparison

Existing work falls short of this design in three recurring ways: fixed sample size, single-method or narrow-comparator scope, and qualitative rather than quantitative stability judgments. Ning et al. [5] test ShapleyVIC's stability at two subsample points (n=2,500, n=500) in an OHCA survival cohort, finding rankings largely reproduced against tree-ensemble baselines, but stop at n=500 rather than N=250-50, compare only a Shapley-ensemble variant against tree-based baselines, and measure agreement qualitatively rather than via a fixed-K failure rate. Alvarez-Melis and Jaakkola [4] show SHAP and LIME can violate local-

robustness, and Lundberg et al. [7] show SHAP dependence plots surface structure PDP misses, both at a single fixed N; two CKD-specific studies [8,9] apply SHAP and LIME jointly to single fixed-size cohorts, judging agreement by clinical plausibility rather than a quantitative metric. SHAP [10] and LIME [11] each have documented failure modes independent of sample size, and neither's original formulation addresses shrinking samples. Strobl et al. [12] show permutation importance over-attributes importance to correlated predictors in random-forest ensembles, a tree-based mechanism not directly applicable here, where multicollinearity instead destabilizes both coefficient- and permutation-based importance through variance inflation, motivating this project's exclusion of BMXWAIST (Section 3.1). Riley and Collins [6] establish sample-size-ladder testing for clinical prediction models generally (n=50-5,000, matching this project's outer bounds), but for model-level stability, not interpretability rankings. Yuan et al. [13] show SHAP's own output is sensitive to background-sample size in a comparable clinical dataset, but manipulate background size rather than training-sample size, with no cross-method comparison. None varies training sample size across a full ladder while holding four methods fixed against an explicit, reproducible failure-rate metric.

Synthesis and Identified Gap

Riley and Collins supply the ladder design; Ning et al. and the method-comparison literature supply multi-method testing; Strobl et al. and Yuan et al. document method-specific instability directly, but none supplies all three jointly, and none defines failure as a fixed, countable, rule-governed event distinguishing insufficient valid draws from dependency-unavailable conditions. This gap motivates Objectives 1 and 2: the multi-method, full-ladder comparison, and the separation of model/draw validity failure from method-level computational failure.

3. Methods

3.1 Data Source and Outcome Definition

Data are drawn from six consecutive cycles of the National Health and Nutrition Examination Survey (NHANES), spanning 2007–2008 through 2017–2018, merged into a single combined dataset of demographic, laboratory, and questionnaire records prior to any exclusion.

This merged dataset was reduced to the final analytic sample through the exclusion steps in Table 1.

Table 1. Sample exclusion criteria and counts

Step	Criterion	Rows remaining	Excluded	Reason
0	Raw merged NHANES 2007–2018	59842	N/A	N/A
1	Age filter	36580	23262	Age < 18 removed
2	Questionnaire-refusal filter	36506	74	DIQ010/BPQ020 coded refusal/don't know (-1)
3	Labs-completeness filter	32298	4208	Missing LBXSCR, URXUMA, or URXUCR
4	Glucose-sentinel filter	32297	1	LBXSGL == 777
5	Predictor-completeness filter	31958	339	Missing value in any of the 7 predictors or has_CKD

The outcome was derived directly from laboratory measurements, not a pre-existing NHANES flag: estimated glomerular filtration rate (eGFR) used the CKD-EPI 2021 race-free creatinine equation [14] (serum creatinine LBXSCR, age RIDAGEYR, sex RIAGENDR). Urine

albumin-to-creatinine ratio (ACR) was calculated from urine albumin (URXUMA) and urine creatinine (URXUCR), in mg/g. Consistent with standard clinical thresholds [15], a participant was classified as CKD-compatible-abnormality positive (has_CKD = 1) if eGFR < 60 mL/min/1.73m² or ACR ≥ 30 mg/g. The three laboratory columns used to derive this outcome (LBXSCR, URXUMA, URXUCR) were excluded from the predictor set throughout, since retaining them would let the model reconstruct its own label. BMXWAIST (waist circumference) was excluded from the seven-predictor set for its high collinearity with BMXBMI ($r = 0.9055$), which inflates coefficient variance and destabilizes permutation-based importance independent of the random-forest-specific mechanism (Section 2.2). The final predictor set comprised age (RIDAGEYR), sex (RIAGENDR), body mass index (BMXBMI), non-fasting serum glucose (LBXSGL, distinct from the fasting measure LBXGLU), diagnosed hypertension (BPQ020), and diabetes status (DIQ010), the last coded as a three-level categorical variable (none, borderline, diagnosed) and represented as two dummy columns against a "none" reference level.

NHANES gives only a single cross-sectional measurement, with no repeat window to confirm the three-month persistence a clinical CKD diagnosis requires [15]; the outcome is therefore described throughout as an NHANES-defined CKD-compatible abnormality rather than a diagnosed case of chronic kidney disease.

NHANES's survey weights, strata, and PSUs support population-representative inference but are not used here: this study evaluates method stability within a fixed, resampled pool, not population prevalence (Section 3.2).

Baseline characteristics of the analytic sample are reported in Supplementary Table S2.

3.2 Sampling Design and Pool Overlap

The exclusion pipeline in Section 3.1 reduced the raw merged NHANES sample from 59,842 to a final analytic pool of 31,958 participants, of whom 5,336 (16.7%) met the CKD-compatible abnormality threshold (Table 1). This pool is the fixed population from which all subsequent ladder draws were taken.

A seven-rung sample-size ladder was constructed at N = 5,000, 2,500, 1,000, 500, 250, 100, and 50, extending precedent for structured sample-size-ladder testing [6] downward to represent an extreme low-resource floor. A fixed K = 100 repeated draws were taken at each rung, so sample size remained the only variable changing between rungs. Each rung used its own disjoint block of 100 seeds, so rungs are independent sets of draws, not nested subsamples.

Draws were taken without replacement, not stratified by outcome, mirroring a real walk-in-clinic setting; this was expected to, and did, produce draws with too few CKD-compatible-abnormality-positive cases for reliable model fitting at smaller rungs, reported as a finding (Section 3.3, Results 4.1) rather than corrected. No redraw policy was applied: a failed draw (Section 3.3) was never replaced, so the reported failure rate at each rung is always exactly failures ÷ 100.

Because the analytic pool (N=31,958) is not much larger than the top rung (N=5,000), draws at the top of the ladder overlap substantially. This was quantified directly across all 4,950 possible draw pairs per rung (Supplementary Table S1).

Overlap (Supplementary Table S1) is a meaningful ceiling-effect concern at N=5,000 only, where draws share nearly one-sixth of their rows on average, falling below 3% from N=1,000 downward and below 1% from N=250 downward.

3.3 Model Validity Checks (Rules 1–3)

Not every draw produced a model that could be reliably fit. Three validity rules were applied to each draw's fitted logistic regression (near-unregularized, C=1e6), in sequence, before any interpretability method was computed. A draw failing any rule was excluded from stability calculations and counted toward a separate model/draw validity failure rate (Results 4.1), distinct from method-specific computational failure (Section 3.4).

Rule 1: Minimum event count (EPV floor). EPV, the number of CKD-compatible-abnormality-positive cases divided by the number of predictors (seven), below 2 (fewer than 14 positive cases) rejected a draw before model fitting. This floor is looser than the classical, not universally supported, EPV≥10 convention [16,17]; EPV=5 and EPV=10 were tested as a sensitivity analysis (Section 3.6).

Rule 2: Convergence check. A draw was rejected if the logistic regression optimizer failed to converge; convergence warnings were captured as the rejection reason, never suppressed.

Rule 3: Coefficient magnitude check. A draw was rejected if any fitted coefficient's absolute magnitude exceeded 10, or exceeded 0.1 with a standard error more than five times that magnitude, or was below 0.1 with a standard error exceeding 3, avoiding misclassification of small, precisely estimated null effects. Standard errors were computed from the observed Fisher information matrix (Wald). These thresholds were developed from this study's own data, not set a priori, so every draw passing Rules 1 and 2 was also refit using Firth's bias-reduced logistic regression [18] as an independent check (Supplementary S1.2).

3.4 Interpretability Methods Compared

Four methods were compared on how consistently each identifies the same important predictors of the CKD-compatible-abnormality outcome across repeated draws at a given sample size.

Two are derived directly from the fitted logistic regression. Standardized regression coefficients (|β|, Section 3.3's fit) served as the primary importance reference, sitting on a common standardized scale unlike raw odds ratios or p-values.

Permutation importance was the in-sample AUC drop from independently shuffling each predictor. AUC was used over accuracy because 16.7% prevalence lets a majority-class classifier reach 83.3% accuracy using no predictor, compressing accuracy-drop into noise; AUC instead has a fixed 0.5 chance baseline and is threshold-independent (Section 3.7). This in-sample choice is tested against a held-out alternative in Section 3.6.

The remaining two are post-hoc, model-agnostic local explanation methods, aggregated to a global ranking. SHAP used shap.LinearExplainer [10], background/masker set to the full draw rather than the library default. LIME used LimeTabularExplainer [11] on a fixed 50-instance seeded subsample per draw, since LIME has no native global-importance mode and full-instance LIME was intractable at N up to 5,000 across 100 draws per rung; this disclosed asymmetry is evaluated in Section 3.6 and Supplementary S1.3/S1.4.

Permutation importance follows Fisher, Rudin, and Dominici's model-agnostic formulation [19], which defines the measure generically for any scoring function and is instantiated here as AUC.

Rules 1–3 (Section 3.3) are applied before any method is computed, excluding the same draws for all four methods at every rung; each method's success or failure is then recorded independently (e.g., a numerical error internal to SHAP or LIME). Results 4.1 reports both outcomes, since they answer different questions.

Full-ranking extraction. Each method's output was recorded as a complete ranking across all seven predictors, rather than a top-5 extraction, since a top-5-only approach would need a shared fallback rank for excluded predictors, invalid for Spearman correlation. Full rankings allow valid Spearman correlation and Kendall's tau (Section 3.5).

3.5 Stability Metrics

Stability was assessed pairwise across all valid draws within a rung, for each method separately, using two complementary measures applied to the full seven-predictor ranking (Section 3.4).

Jaccard similarity was calculated on the set of top-k ranked predictors between each pair of valid draws, averaged across all pairs per method per rung. k = 3, 5, and 7 were all computed; top-5 was retained as the primary value, with top-3/top-7 as a sensitivity check (Section 3.6, Supplementary S1.5). Chance baselines: for two random top-5 subsets of seven predictors, the expected Jaccard similarity is 0.569 (minimum 0.429); for two random full rankings, the expected Spearman correlation is 0.

Rank correlation used full-ranking Spearman correlation between each pair of valid draws' complete seven-predictor rankings. Kendall's tau was computed alongside Spearman on the same full rankings as a stricter concordance check, and both are reported together throughout.

Pairwise Jaccard and rank-correlation calculations were implemented as an exact closed-form matrix computation, validated against a direct pairwise implementation to within 1×10^{-9} on real data.

Uncertainty estimation. Wilson score 95% confidence intervals [20] were used for failure rates and other binomial proportions, since Wilson behaves correctly at boundary values (0.000, 1.000), which this study's failure rates reach at several rungs, where a normal-approximation interval would be degenerate. For Jaccard and rank-correlation metrics, 95% confidence intervals used draw-level bootstrap resampling [21] (2,000 iterations, percentile method), resampling at the draw level to preserve the repeated-draw design's independence.

3.6 Sensitivity Analyses

Seven disclosed design choices (the EPV floor; Rule 3's coefficient-magnitude threshold vs. Firth's estimator; the diabetes-borderline sparse-category collapse; permutation importance's in-sample evaluation; the LIME subsample size; the top-5 Jaccard cutoff; and whether ranking reproducibility reflects correctness against a full-pool reference) were each tested against an alternative or independent check; full methodology and results are in Supplementary Material S1.

3.7 Predictive Performance Evaluation

The stratified 80/20 train/test split (Supplementary S1.4) evaluated held-out predictive performance across the ladder (Objective 5), gated independently of Rules 1–3: split-validity (training EPV ≥ 2, test positive count ≥ 5) is assessed separately from model-validity status.

For every split-valid draw, a logistic regression was fit on the training portion only (C=1e6), with the scaler fit on training and applied to test without refitting. Discrimination used AUC

[22] and calibration used Brier score [23], both computed at every valid draw and rung with 95% bootstrap CIs (2,000 iterations, percentile method, Section 3.5).

A calibration curve was not attempted: held-out test portions at smaller rungs contain as few as five positive cases by design, insufficient for a reliable curve; this is a disclosed limitation, not an omission.

3.8 Formal Statistical Comparison Between Methods

Method-ranking comparisons in Results 4.2 and Discussion are descriptive unless accompanied by a formal test of whether an observed stability difference between two methods is distinguishable from sampling variability.

For every pair of the four methods, at every rung except N=50 (no valid draws, Results 4.1), a paired draw-level bootstrap test compared stability: at each iteration, the same resampled draw indices (with replacement) were applied to both methods, preserving the paired structure since both were computed on the same original draws. The difference in stability (method A minus method B) was recorded per iteration, yielding a 95% confidence interval and two-sided p-value. With 2,000 iterations, comparisons in which no iteration crossed zero are reported as $p<0.0005$.

This was applied to all six method pairs at six rungs (N=50 excluded) across two stability metrics (Spearman, Jaccard@5), for 72 comparisons total; the Benjamini-Hochberg procedure controlled the false discovery rate at $q \leq 0.05$ across all 72 jointly [24], rather than treating each raw p-value as independently interpretable.

3.9 Predictor-Level Analysis

This predictor-level analysis identifies which individual predictors were most vulnerable to instability as sample size decreased (Objective 4). For each predictor, a top-5 appearance rate was calculated as the proportion of valid draws, at a given rung and method, in which that predictor appeared within the top five of the full seven-predictor ranking (Section 3.4), averaged across methods to a single per-predictor, per-rung rate reported as a heatmap (Results 4.4), reusing the full rankings from Section 3.5 at all six rungs with valid draws.

3.10 Software and Reproducibility

Implemented in Python 3.14.6: pandas 3.0.5, numpy 2.4.6, scikit-learn 1.9.0, scipy 1.18.0, shap 0.52.0, lime 0.2.0.1, firthmodels 0.8.1 (Rule 3 comparator, Section 3.6). All draws used fixed, deterministic seeds for exact reproducibility. Analysis code, dataset-construction scripts, and all output tables are available at https://github.com/AbdulazeezHaneef/ckd-interpretability-data-scarcity, archived at https://doi.org/10.5281/zenodo.23071370.

Generative AI (Claude, Anthropic) was used to assist with writing and debugging the statistical analysis code used to generate the study's results, and separately to assist with condensing, restructuring, and checking the manuscript text for submission. After using this tool, the authors reviewed and edited all content and take full responsibility for the content of the published article.

4. Results

4.1 Model/Draw Validity Across the Sample-Size Ladder

No method-level computational failure occurred at any evaluable rung, for any method (every cell 0.0%, or n/a at N=50 where no valid model existed). The stability degradation in Section 4.2 therefore does not reflect computational failure.

Model/draw validity failure (a draw rejected by Rule 1, 2, or 3) rose sharply as sample size decreased.

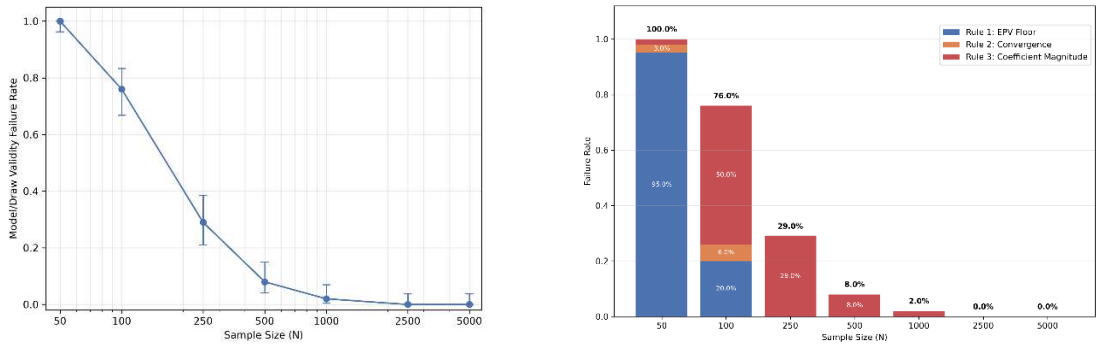


Figure 1: Model/Draw Validity Failure Rate

Figure 1(a) reports the overall failure rate at each rung with 95% confidence intervals; Figure 1(b) decomposes it by responsible rule as point estimates only, since per-component confidence intervals are not well-defined for stacked proportions.

Failure rose monotonically from 0.0% (N≥2,500) to 100.0% (N=50), reflecting low event counts and sparse-category separation, with the two smallest rungs dominated by different rules: at N=100, Rule 3 accounts for 50.0 of the 76.0 points (EPV floor 20.0, convergence 6.0); at N=50, the EPV floor dominates (95.0 of 100.0). Mean positive cases per draw were 42.3 at N=250 (minimum 28), 17.2 at N=100 (minimum 5) and 8.2 at N=50 (minimum 4), against the 14-case floor. From N=1,000 to N=250, every validity failure was a Rule 3 failure on the

diabetes-borderline dummy (2.3% prevalence): 2 of 2 at N=1,000, 8 of 8 at N=500, 26 of 29 at N=250, 44 of 50 at N=100, so the failure curve reflects this design's rare category as well as sample size.

Two independent sensitivity checks confirm this: collapsing the two diabetes dummies into a single binary predictor cut failure from 2% to 0% at N=1,000, 8% to 0% at N=500, 29% to 0% at N=250 and 76% to 20% at N=100; gating Rule 3 on Firth's bias-reduced fit instead of MLE produced a comparable drop (29%→3% at N=250, 76%→37% at N=100), with 27 and 43 draws respectively becoming newly valid against 1 and 4 moving the other way (Supplementary S1.6).

4.2 Interpretability Stability Across the Ladder

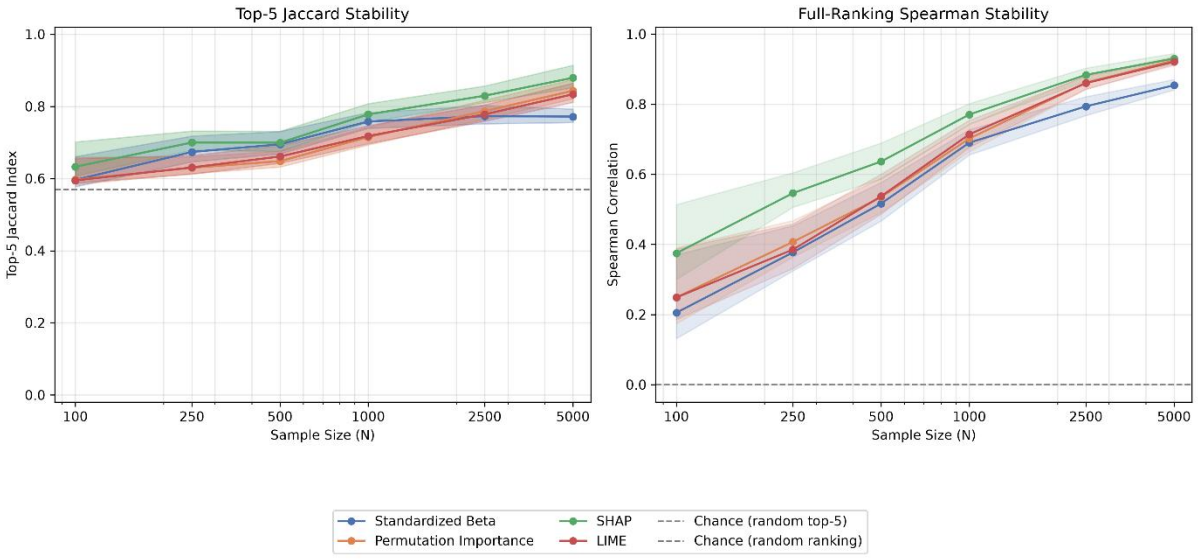


Figure 2: Stability Degradation

Figure 2 reports top-5 Jaccard similarity and full-ranking Spearman correlation (Section 3.5) for all four methods across the ladder, with bootstrap 95% CI bands. Kendall's tau, computed alongside Spearman, showed the same degradation pattern at every rung and is not plotted separately.

All four methods declined as N decreased, but not at the same rate. Standardized-β had the lowest full-ranking Spearman at every rung, falling from 0.854 (N=5,000) to 0.205 (N=100); on top-5 Jaccard it was lowest only at the two largest rungs (0.772, 0.773), falling to 0.596, after which permutation importance or LIME was lowest. On Spearman it started lowest rather than falling fastest: its decline (0.649) was smaller than permutation importance's (0.676) or LIME's (0.672), while SHAP had both the highest values throughout and the smallest decline (0.555). At N=50, no method produced a valid model (Section 4.1), so no stability estimate exists there.

Confidence intervals widened substantially as N decreased; overlapping intervals at several rungs (particularly N=100) are addressed formally in Section 4.3, not asserted from point estimates alone.

Against chance (Section 3.5), top-5 Jaccard at N=100 was 0.595–0.597 for standardized-β, permutation importance and LIME (lower bounds 0.577–0.588), within ~0.03 of the 0.569 chance value; SHAP was 0.633 (95% CI 0.607–0.701). Spearman at N=100 (0.205–0.375; lower bounds ≥0.131) stayed above the chance value of 0.

4.3 Formal Statistical Comparison Between Methods

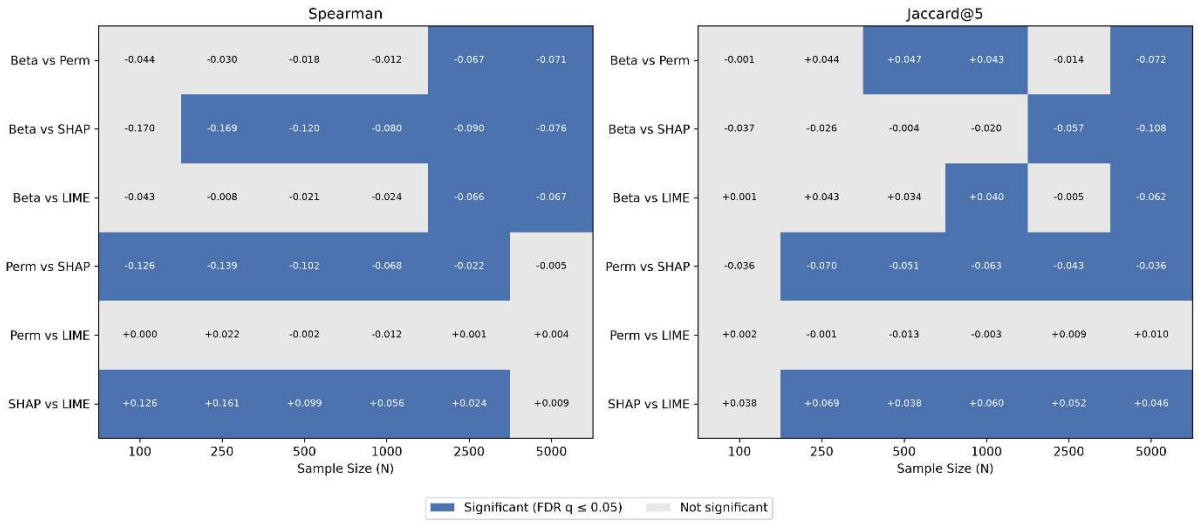


Figure 3: Significance Heatmap

Figure 3 reports the paired draw-level bootstrap comparisons (Section 3.8), FDR-corrected across all 72 comparisons jointly. Each cell reports the point-estimate difference in stability between the two methods named in its row (first minus second); shading indicates significance at FDR $q \leq 0.05$. 36 of 72 were significant.

SHAP was the most consistently distinguishable method: significant against permutation importance and LIME on 10 of 12 comparisons each, across nearly every rung and both metrics, and against standardized-β on Spearman at five of six rungs but only the two largest rungs (N=5,000, 2,500) on Jaccard@5 (7 of 12 overall). Standardized-β differed significantly from permutation importance on 5 of 12 and from LIME on 4 of 12.

Permutation importance and LIME were not significantly distinguishable from each other at any rung or metric (0 of 12).

N=100 is where significance broke down most sharply: only 2 of 12 comparisons remained significant (0 of 6 Jaccard@5, 2 of 6 Spearman: permutation-SHAP and SHAP-LIME), versus 8, 8, 7, 6 and 5 of 12 at N=5,000 through N=250. Any claim that one method is more stable should specify sample size, method pair, and metric. Within draws, the methods form two groups: at N=5,000, mean pairwise Spearman agreement was 0.98 for permutation–LIME, 0.95 for permutation–SHAP and 0.96 for SHAP–LIME, against 0.83–0.85 between standardized-β and the other three; β's agreement fell to 0.53–0.62 at N=250, while permutation–LIME remained 0.88 at N=100. The absent permutation–LIME difference thus reflects near-identical rankings, not established equivalence.

4.4 Predictor-Level Vulnerability

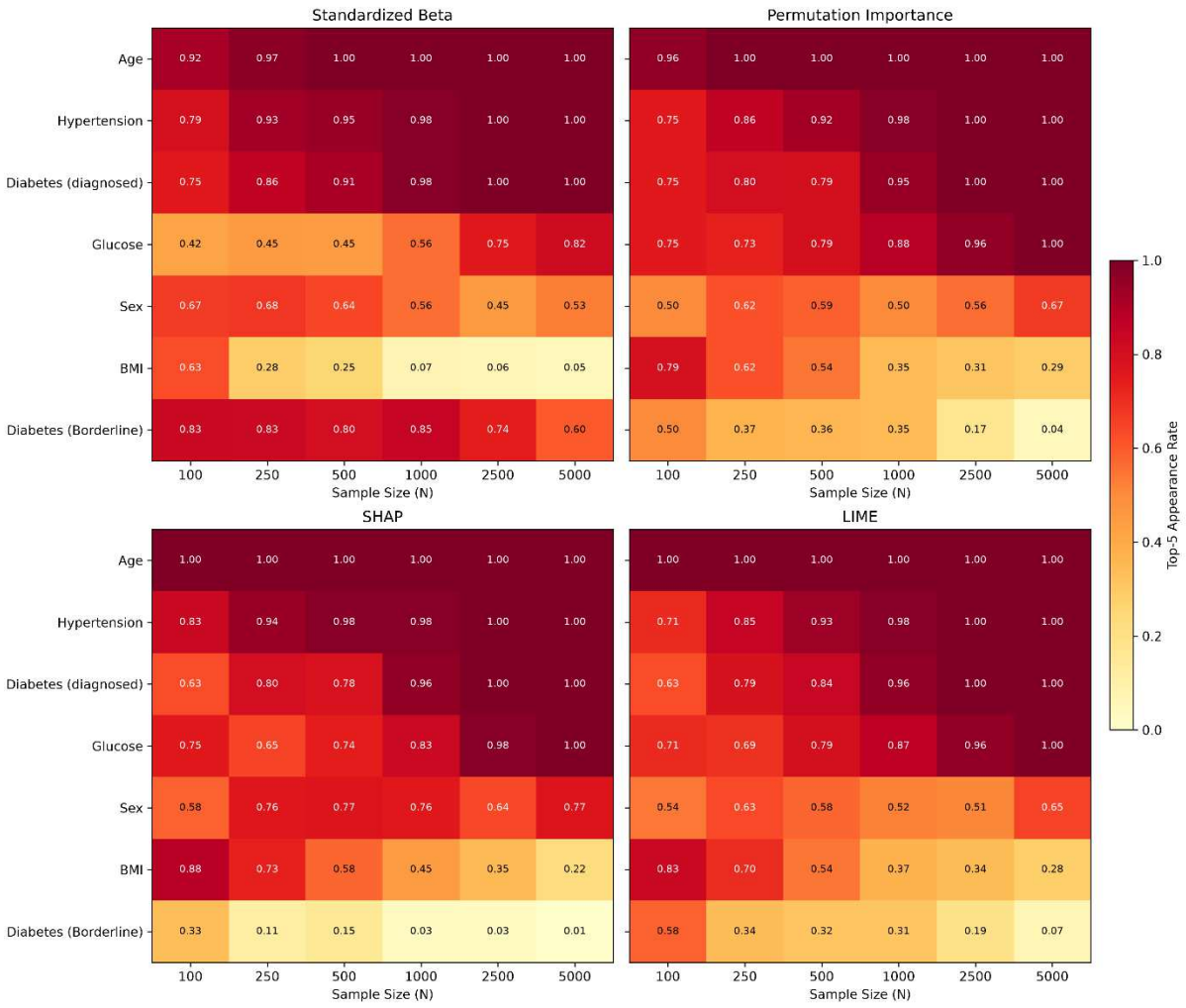


Figure 4: Predictor Heatmap

Figure 4 reports the top-5 appearance rate for each of the seven predictors, averaged across all four methods, at each of the six rungs with valid draws (N=50 excluded). Predictors are ordered by their N=5,000 baseline appearance rate, descending.

Age was the most stable predictor, near-ceiling (96.9–100.0%) at every rung. Diabetes-yes and hypertension started similarly high at N=5,000 but declined further by N=100; glucose declined more steadily; sex stayed flat and moderate throughout (Figure 4).

BMI and diabetes-borderline had the lowest top-5 rates at N=5,000 (21.0% and 18.0%), and each behaves differently as N falls.

BMI rose from 21.0% (N=5,000) to 78.1% (N=100), opposite the stronger predictors, moving past the chance baseline for top-5 inclusion among seven predictors (5/7 ≈ 71.4%) at N=100, where its rate under permutation importance, SHAP and LIME was 79%, 88% and 83%. This matches established resampling-based variable-selection behavior: strong-effect predictors are selected with high probability across replications, while weak-effect predictors' selection frequency drifts toward chance as estimation noise grows relative to a thin true signal [25,26].

Diabetes-borderline showed a different, method-specific pattern: consistently low at large N (18.0% at N=5,000, 28.3% at N=2,500) before rising to 56.3% by N=100, never converging to age's near-ceiling stability. Collinearity with diabetes-yes was ruled out ($r = -0.057$, VIF = 1.03). Its top-5 rate at N=5,000 was 60% under standardized-β, 4% under permutation importance, 1% under SHAP and 7% under LIME; at N=100: 83%, 50%, 33% and 58%. Standardized-β's mean rank for the predictor moved from 5.06 to 2.92 as N fell, against 6.87 to 5.92 for SHAP; SHAP excluded it from the top five in 99% of N=5,000 draws, which is stable, not unstable (it is present in only 2.28% of the analytic sample, 12.36% for diabetes-yes). A high standardized-β rank for so rare a predictor is consistent with sparse-cell coefficient inflation, the same behaviour that dominated Rule 3 failures (Section 4.1); the averaged rate in Figure 4 conflates these method-specific patterns.

A full-pool (N=31,958) fit confirms the direction of this split: standardized-β ranks diabetes-borderline above BMI, while the other three methods rank it below BMI (Supplementary S1.6).

4.5 Predictive Performance and Its Relationship to Stability

Table 2 reports split-validity by rung (Section 3.7); Figure 5 reports held-out AUC alongside mean interpretability stability (Spearman, averaged across methods) on different y-axis ranges (AUC: 0.45–0.85; stability: 0.0–1.0) to keep each series legible; the two should not be compared by slope.

Table 2. Split-validity by rung

N	Split-valid (n)	Train-EPV fail (n)	Test-positives fail (n)	Split-valid rate
5000	100	0	0	100.0%
2500	100	0	0	100.0%
1000	100	0	0	100.0%
500	100	0	0	100.0%
250	100	0	0	100.0%
100	12	45	43	12.0%
50	0	100	0	0.0%

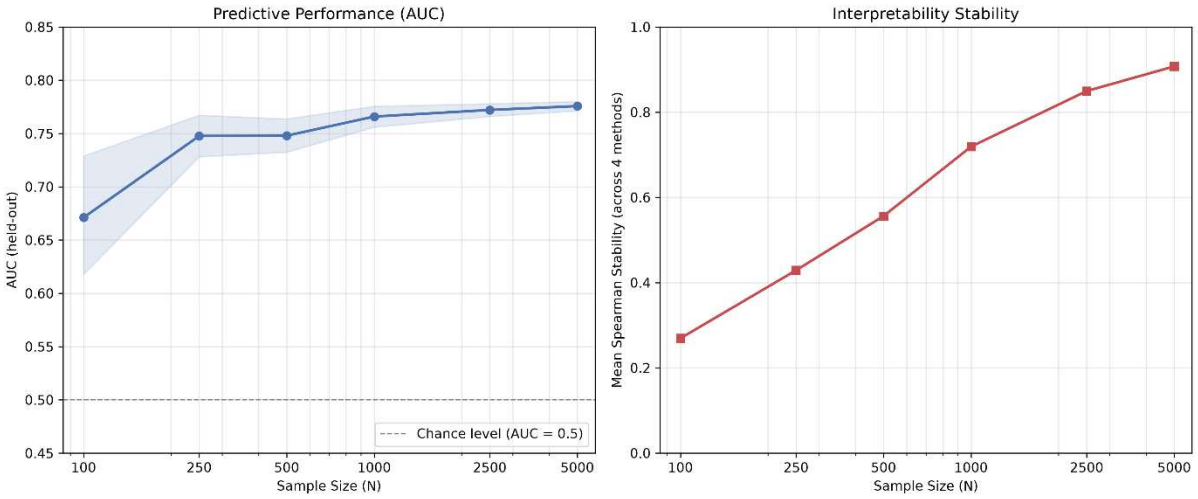


Figure 5: Performance Stability Decoupling

Held-out performance and interpretability stability followed markedly different trajectories. AUC declined only modestly, from 0.776 (N=5,000) to 0.748 (N=250); mean stability declined far more sharply over the same range, from 0.907 to 0.429. Brier score showed a similarly modest shift (0.115→0.124). Across N=5,000 to N=250, AUC and Brier stayed close to their N=5,000 values even as mean stability fell by more than half.

Performance did not remain robust indefinitely: at N=100, split-validity dropped sharply (12.0% vs. 100.0% at N=250 and above) and AUC among split-valid draws fell to 0.671, with a substantially wider CI. Split-validity (Table 2) and model-validity (Rules 1–3) are gated on different criteria and need not track together: at N=100 only 4 of 24 model-valid draws also meet split-validity, while 8 of 12 split-valid draws fail model-validity (Supplementary S1.4), consistent with N=100 being where model validity, split validity, and method-distinguishability all break down together (4.1, 4.3). At N=50, no draw met split-validity, so no held-out estimate exists there.

The central pattern is a decoupling between predictive performance and interpretability stability: performance remains comparatively stable across most of the ladder even as the explanations underlying that same model become substantially less reproducible.

5. Discussion

5.1 Model Validity Is Not Method-Specific

The uniform 0.0% computational failure rate across all four methods (Section 4.1) is a property of the single upstream validity gate they sit behind (Rules 1–3, Section 3.3), not of the methods themselves. Separating model/draw validity from method-specific failure avoids the

misleading conclusion that SHAP or LIME becomes computationally unreliable as sample size shrinks, when in fact all four compute without error whenever given a model to work from.

5.2 Interpretability Stability Degradation

The degradation pattern in Section 4.2 is this study's core empirical finding, directly answering Objective 1; Kendall's tau, computed alongside Spearman as a second check, shows the same downward trajectory at every rung, so the degradation is not an artifact of one measure. Two caveats apply: confidence intervals widen substantially as N decreases, so a point estimate at N=100 carries meaningfully more uncertainty than one at N=5,000; and at N=50, no method produced a valid model at all (Section 4.1), so this degradation trend runs from N=5,000 to N=100, not the full nominal ladder.

A six-predictor variant collapsing the two diabetes dummies into one did produce valid models at N=50 (10 of 100 draws), giving a first, fragile stability estimate there: mean Spearman 0.265, consistent with continued degradation rather than a floor (Supplementary S1.6).

5.3 Formal Significance of Method Rankings

Section 4.3's paired bootstrap comparison tests whether these descriptive stability differences (Section 5.2) are statistically distinguishable from sampling variability: SHAP was the most consistently distinguishable method, while permutation importance and LIME were never significantly distinguishable from each other. No mechanism for SHAP's comparative edge is proposed here; establishing one would need a dedicated analysis beyond this manuscript's scope.

These findings are not a fixed, universal ranking. N=100 is where significance broke down most sharply: 0 of 6 Jaccard@5 comparisons and 2 of 6 Spearman comparisons were significant there, against 8, 8, 7, 6 and 5 of 12 comparisons at N=5,000 through N=250.

5.4 Predictor-Level Vulnerability

Objective 4 asks which individual predictors are most vulnerable to instability as sample size decreases (Section 4.4, all six usable rungs). At N=5,000, Age, diagnosed diabetes and hypertension all sit near ceiling and appear similarly stable, but only Age holds that stability to the smallest usable rung; the other two degrade meaningfully, just more slowly than diabetes-borderline.

BMI and diabetes-borderline behave differently: BMI's top-5 rate rose past the chance baseline, consistent with a weak signal whose selection frequency drifts toward chance as estimation noise grows [25,26]; diabetes-borderline's pattern is method-specific (Section 4.4), with collinearity with diabetes-yes ruled out, and is consistent with sparse-category behaviour (2.28% prevalence), though not established as its cause.

5.5 Performance and Stability Decouple

Section 4.5's finding directly answers Objective 5: predictive performance held comparatively steady across most of the ladder while interpretability stability declined far more sharply. Part of this gap is mechanical (full-ranking Spearman over seven items is inherently higher-variance than an aggregate scalar like AUC), but a model built at N=250 still discriminates nearly as well as one at N=5,000 while relying on a materially different, less reproducible set of top predictors.

Within rungs, a draw's held-out AUC showed no consistent relationship with how closely its ranking matched a full-pool reference ranking (Pearson correlations −0.24 to +0.17 across methods and rungs, no consistent sign; Supplementary S1.6): a well-discriminating draw's feature ranking is not more likely to be correct.

The decoupling is not uniform: the same N=100 breaking point reappears here in both split-validity and AUC, so performance does eventually degrade alongside stability, just much later on the ladder. This gap, where the decoupling is most consequential, has not previously been examined jointly in prior CKD-focused explainable-AI work (Section 5.6).

Confidence intervals widen mechanically at low N, so validity failure (5.1), split-validity collapse, and loss of formal significance (5.3) all break down together at N=100: this study's floor for confident inference, distinct from N=50's floor for any inference at all.

5.6 Comparison with Existing Literature

Prior CKD-focused explainable-ML work establishes SHAP and LIME can be computed on clinical data and surface plausible risk predictors [8,9], and that interpretability methods can themselves be unstable, independent of model performance, has precedent outside CKD [4]. To our knowledge, no prior CKD-focused study evaluates interpretability-method stability across a sample-size ladder, tests between-method significance, or examines performance and stability jointly; this manuscript does so using model-derived and post-hoc methods side by side (Section 3.4, 5.5).

Explainer choice limits direct comparison with prior CKD-specific SHAP work, which generally uses tree-based TreeExplainer [7] rather than this manuscript's shap.LinearExplainer (Section 3.4); rankings are therefore not directly comparable, though the stability question is unaffected.

5.7 Practical Implications

Data scarcity here is simulated by subsampling a large national survey, not observed in a low-resource clinical dataset; these implications describe how sample size interacts with stability under the conditions tested, not a deployment recommendation.

A model built from 250 cases discriminates nearly as well as one from 5,000 (Section 5.5), while its apparent top predictors may vary considerably between similarly-sized samples, so caution should scale with sample size rather than performance. These thresholds depend on this study's 16.7% prevalence and seven predictors, and should be restated as expected event counts for other settings: a 150-patient pilot at the same prevalence would expect about 25 events, below the N=250 rung, where mean stability was already 0.38–0.55.

Two findings offer partial mitigation. Method choice matters: SHAP was the most stable and distinguishable method (Section 5.3), while standardized-β, this study's own primary reference, had the lowest Spearman stability at every rung; since all four derive from the same fitted model, this is not evidence SHAP is more accurate, and no method should be chosen on stability alone. Predictors also do not degrade equally: age stayed reliably top-ranked (Section 5.4), while BMI and diabetes-borderline, the two least often top-ranked at N=5,000, changed most as sample size fell, in different ways (Section 4.4).

5.8 Limitations

Age and sex enter the CKD-EPI equation that defines the outcome, so their importance is partly built into the label; age's stability (Section 4.4) should be read accordingly.

The Firth-gated sensitivity check (Section 4.1, Supplementary S1.6) confirms this: the primary analysis's valid-draw set at N=250 and N=100 already excludes a share of draws a less conservative but defensible standard would keep.

Findings describe an NHANES-defined CKD-compatible abnormality from a single cross-sectional measurement, not a diagnosed condition, and method stability within a fixed analytic pool, not population estimates, because survey weights, strata, and PSUs were not used (Section 3.1).

Of 59,842 raw records, 31,958 (53.4%) remained after exclusions (Table 1), mostly from the age filter, so this sample may differ from the full population. The predictor-completeness filter also excluded missing cases before any ladder draw, so every rung draws from an already-complete-case pool; real-world low-resource data carries more missingness at smaller samples, so real-world instability is plausibly worse than shown here.

BMXWAIST was excluded for collinearity with BMXBMI (r=0.9055, Section 3.1); BMI's apparent importance may partly reflect a merged BMI/waist-circumference signal.

Calibration curves were not attempted (Section 3.7) because held-out test portions at smaller rungs contain as few as five positive cases.

This study uses a single near-unregularized logistic regression (Section 3.3) rather than a tree ensemble or other high-capacity model; that choice suits data scarcity and makes model-derived coefficients usable as one of the four compared methods, but findings should not be assumed to generalize beyond it (Section 5.6).

Stability at N=250 and N=100 is computed only on draws that passed Rules 1–3 (71 and 24 of 100 draws). Valid draws are likely to contain more events, so small-N stability may be overstated, and rungs compare differently selected populations.

Rule 3's thresholds were developed from this study's own data (Supplementary S1.2); failures from N=1,000 to N=250 almost always involved the diabetes-borderline coefficient (2.3% prevalence), so the validity-failure curve reflects sparse-category separation as well as sample size.

Agreement between draws measures reproducibility, not correctness: a method can be stable and wrong.

This analysis also draws on a single, US-based data source; prevalence, predictor distributions, and therefore the specific sample sizes at which instability sets in will differ in other populations, including the low-resource settings this work is partly motivated by.

Permutation importance's held-out check (Supplementary S1.4) becomes unreliable at N=100, where only 4 of 24 model-valid draws met split-validity; its strong in-sample/held-out agreement (0.63–0.95, N=5,000–250) should not extend there.

6. Conclusions

Across the ladder, predictor-ranking stability fell sharply (mean Spearman 0.907 at N=5,000 to 0.429 at N=250) while held-out discrimination changed little (AUC 0.776 to 0.748). Method computation never failed; validity failures arose upstream in the model-fitting rules, mostly Rule 3 failures on the rare diabetes-borderline category from N=1,000 to N=250. Standardized coefficients had the lowest Spearman stability at every rung and SHAP the highest, though all four derive from one fitted model, and rank instability concentrated in weakly and sparsely represented predictors while age remained stable.

Four lessons follow for small-sample clinical prediction work: report the expected number of outcome events and the model-validity failure rate alongside performance; report ranking stability beside AUC, with its chance baseline (Jaccard@5 0.569, Spearman 0); do not read the rank of weak or rare predictors from small samples; and, for linear models, method choice matters mainly through how each method scales predictors, since standardized-β and SHAP are not on the same scale and a full-pool reference fit shows this changes which predictors each method ranks highest, not only how stable the ranking is.

These findings come from simulated data scarcity in one dataset and should be tested on observed small clinical samples before informing practice.

List of Abbreviations

ACR: Albumin-to-Creatinine Ratio

AUC: Area Under the (Receiver Operating Characteristic) Curve

BMI: Body Mass Index

CI: Confidence Interval

CKD: Chronic Kidney Disease

eGFR: estimated Glomerular Filtration Rate

EPV: Events Per Variable

FDR: False Discovery Rate

KDIGO: Kidney Disease: Improving Global Outcomes

LIME: Local Interpretable Model-agnostic Explanations

NHANES: National Health and Nutrition Examination Survey

SHAP: SHapley Additive exPlanations

VIF: Variance Inflation Factor

Declarations

Clinical trial number

Clinical trial number: not applicable.

Ethics approval and consent to participate

Ethics statement: This study is a secondary analysis of publicly released, de-identified NHANES data. NHANES protocols were approved by the National Center for Health Statistics Research Ethics Review Board with written informed consent; no participants were contacted and no identifiable data were accessed, so no further approval was required.

Consent for publication

Not applicable.

Availability of data and materials

This study used publicly available data from the National Health and Nutrition Examination Survey (NHANES), accessible at https://www.cdc.gov/nchs/nhanes/. No new data were generated for this study. Analysis code and derived output tables are available at https://github.com/AbdulazeezHaneef/ckd-interpretability-data-scarcity, archived at https://doi.org/10.5281/zenodo.23071370 [27].

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

HA conceptualized the study, developed the methodology and software, performed the formal analysis, investigation and data curation, produced the visualizations, led project administration, and drafted and revised the manuscript. JOB and AAI supervised the work, contributed to the methodology, and assisted with project administration. REA contributed to the methodology and reviewed and edited the manuscript. AM-Y supervised the work. KEL contributed to conceptualization, supervision and methodology, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Author	Roles
Abdulazeez	Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing: original draft, Writing: review & editing, Visualization, Project administration
Bello	Supervision, Methodology, Project administration
Issa	Supervision, Methodology, Project administration
Eyitayo	Methodology, Writing: review & editing
Musa	Supervision
Yusuf	
Lasisi	Conceptualization, Supervision, Methodology, Writing: review & editing

Acknowledgements

Not applicable.

Additional files

File name: Additional file 1.

File format: .docx

Title of data: Supplementary Material.

Description of data: Baseline characteristics and pairwise row-overlap table (Supplementary Tables S1–S2); four supplementary figures (S1–S4) showing EPV sensitivity, Firth divergence, sensitivity-method agreement, and top-k Jaccard comparison; and six sensitivity analyses (S1.1–S1.6) testing the EPV threshold, Rule 3 validation, LIME subsampling, permutation importance held-out evaluation, top-k choice, and Rule 3 sparse-category/reference-ranking robustness.

References

1. Deng L, Guo S, Liu Y, Zhou Y, Liu Y, Zheng X, Yu X, Shuai P. Global, regional, and national burden of chronic kidney disease and its underlying etiologies from 1990 to 2021: a systematic analysis for the Global Burden of Disease Study 2021. BMC Public Health. 2025 Feb 17;25(1):636.
2. Ulasi II, Awobusuyi O, Nayak S, Ramachandran R, Musso CG, Depine SA, Aroca-Martinez G, Solarin AU, Onuigbo M, Luyckx VA, Ijoma CK. Chronic kidney disease burden in low-resource settings: regional perspectives. In Seminars in Nephrology 2022 Sep 1 (Vol. 42, No. 5, p. 151336). WB Saunders.
3. Stanifer JW, Jing B, Tolan S, Helmke N, Mukerjee R, Naicker S, Patel U. The epidemiology of chronic kidney disease in sub-Saharan Africa: a systematic review and meta-analysis. The Lancet Global Health. 2014 Mar 1;2(3):e174-81.
4. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049. 2018 Jun 21.

5. Ning Y, Li S, Ng YY, Chia MY, Gan HN, Tiah L, Mao DR, Ng WM, Leong BS, Doctor N, Ong ME. Variable importance analysis with interpretable machine learning for fair risk prediction. PLOS Digital Health. 2024 Jul 12;3(7):e0000542.
6. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical Journal. 2023 Dec;65(8):2200302.
7. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020 Jan;2(1):56-67.
8. Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports. 2024 Feb 14;14(1):3687.
9. Mehdi Chouit E, Rachdi M, Bellafkih M, Raouyane B. Interpretable machine learning for chronic kidney disease prediction: Insights from SHAP and LIME analyses. Plos one. 2026 Feb 26;21(2):e0343205.
10. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
11. Ribeiro MT, Singh S, Guestrin C. " Why should I trust you?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144).
12. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics. 2007 Jan 25;8(1):25.
13. Yuan H, Liu M, Kang L, Miao C, Wu Y. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. arXiv preprint arXiv:2204.11351. 2022 Apr 24.

686 14. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, Crews DC, Doria A,
687 Estrella MM, Froissart M, Grams ME. New creatinine-and cystatin C–based equations
688 to estimate GFR without race. New England Journal of Medicine. 2021 Nov
689 4;385(19):1737-49.
690 15. Levin A, Stevens PE, Bilous RW, Coresh J, De Francisco AL, De Jong PE, Griffith
691 KE, Hemmelgarn BR, Iseki K, Lamb EJ, Levey AS. Kidney Disease: Improving Global
692 Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for
693 the evaluation and management of chronic kidney disease. Kidney international
694 supplements. 2013 Jan 1;3(1):1-50.
695 16. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the
696 number of events per variable in logistic regression analysis. Journal of clinical
697 epidemiology. 1996 Dec 1;49(12):1373-9.
698 17. Van Smeden M, De Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ,
699 Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic
700 regression analysis. BMC medical research methodology. 2016 Nov 24;16(1):163.
701 18. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993 Mar 1:27-
702 38.
703 19. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning
704 a variable's importance by studying an entire class of prediction models simultaneously.
705 Journal of machine learning research. 2019;20(177):1-81.
706 20. Wilson EB. Probable inference, the law of succession, and statistical inference. Journal
707 of the American Statistical Association. 1927 Jun 1;22(158):209-12.
708 21. Efron B. Bootstrap methods: another look at the jackknife. InBreakthroughs in
709 statistics: Methodology and distribution 1992 Dec (pp. 569-593). New York, NY:
710 Springer New York.

22. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29-36.
23. Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review. 1950 Jan;78(1):1-3.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995 Jan;57(1):289-300.
25. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Statistics in Medicine. 1989;8(7):771-783.
26. Chen CH, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. Statistics in Medicine. 1985;4(1):39-46.
27. Abdulazeez H. ckd-interpretability-data-scarcity (v1.0.0) [software]. Zenodo; 2026. https://doi.org/10.5281/zenodo.23071370.

Supplementary Files

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

- SupplementaryMaterial.docx