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.

RARE-CROSS: development and evaluation of an operational framework for trustworthy rare-event modelling in cross- sectional health surveys

Wingston Felix Ng'ambi

wingston.ngambi@gmail.com

Kamuzu University of Health Sciences

Cosmas Zyambo

University of Zambia

Lawrence Kazembe

University of Namibia

Research Article

Keywords: rare events, class imbalance, cross-sectional survey, prediction modelling, calibration, SMOTE, complex survey, missing data, transportability, reproducibility

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11096153/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

Rare-event analyses in cross-sectional health surveys are often treated primarily as class-imbalance problems, although validity can be compromised by outcome definition, temporal ambiguity, detection-sensitive predictors, complex sampling, missingness, data leakage, probability distortion after balancing and poor transportability. Existing standards address many of these threats separately. We developed RARE-CROSS to integrate them into an operational and auditable workflow.

Methods

RARE-CROSS comprises nine domains organised around DEFINE, PROTECT and PROVE. Development mapped recurrent failure modes to established guidance and translated them into operational guardrails, minimum reporting items and enhanced options. The framework was cross-walked against TRIPOD + AI, PROBAST + AI, PRICSSA, simulation-study and external-validation guidance. Targeted Monte Carlo experiments examined prevalence effects, SMOTE leakage, altered class priors, detection-related predictors and outcome-dependent missingness. An empirical demonstration used 52,572 adults, 2,778 self-reported cardiovascular disease events and 11 African WHO STEPS survey-country datasets from 2014 to 2019.

Results

Mean ROC AUC changed little as simulated prevalence varied from 1% to 10% (0.793 to 0.773), whereas PR AUC increased from 0.105 to 0.328. Correctly nested SMOTE produced PR AUC 0.166, but pre-split SMOTE produced an artefactual 0.799. Prior-prevalence correction reduced the Brier score from 0.178 to 0.023 without changing ranking. A detection-related covariate increased PR AUC from 0.224 to 0.327. In the STEPS demonstration, pooled leave-one-country-out PR AUC was 0.079, with substantial country heterogeneity.

Conclusions

RARE-CROSS provides an evaluated, algorithm-agnostic architecture for designing, auditing and reporting rare-event analyses in cross-sectional surveys. It integrates prevalence-aware performance, calibration, leakage-safe preprocessing, missingness structure, detection-dependence stress testing, survey design, uncertainty and performance beyond development data. The framework is intended to complement, not replace, established reporting and risk-of-bias standards.

Background

Rare outcomes are common targets in population health surveys, including prior myocardial infarction, stroke, severe mental-health outcomes, maternal complications, injuries, rare infections and disabling conditions. Their analysis is statistically difficult because low prevalence interacts with finite event counts, model complexity, sampling structure and outcome ascertainment. Rare-event and imbalanced-learning methods address important parts of this problem, but low prevalence alone does not determine whether a model is valid or useful [1–6].

Rare-event methods span bias-corrected logistic regression, cost-sensitive learning, resampling, penalisation and ensemble approaches [1–4, 7, 8]. Sound regression analysis additionally requires explicit outcome coding, defensible predictor specification, assessment of information relative to model complexity and transparent diagnostics [9]. None of these choices by itself establishes that the target is correctly defined, predictors are temporally plausible, preprocessing is leakage-safe, missingness is handled appropriately, population probabilities remain calibrated, or performance travels to a new survey setting. RARE-CROSS therefore treats imbalance handling as one component of a chained validity problem rather than the organising principle for the entire analysis.

Performance evaluation must also be prevalence-aware. ROC AUC can remain apparently stable while precision among positive predictions deteriorates sharply as prevalence falls; precision-recall analysis therefore provides essential complementary information in rare-event settings [10, 11]. Probability accuracy requires separate assessment: the Brier score, calibration intercept and slope and calibration plots address questions that ranking metrics cannot [12–16]. Where predictions are intended to trigger action, decision-analytic measures such as net benefit provide an additional layer of evaluation [17, 18].

Cross-sectional surveys introduce further complications. Predictors and outcomes are often measured at the same visit, creating temporal ambiguity and the possibility that clinically ascertained predictors encode healthcare contact after the event. Complex survey weights, strata and clusters affect population representation, while missingness may reflect participant nonresponse, derived-variable failure or structural non-collection in an entire country or wave. These mechanisms require different analytical responses and should not be collapsed into a single generic missing-data problem [19–26].

Current methodological standards support many of these safeguards but address different parts of the problem. TRIPOD + AI and TRIPOD emphasise transparent reporting of prediction models; PROBAST + AI and PROBAST focus on risk of bias and applicability; PRICSSA addresses complex survey analysis; and contemporary guidance distinguishes internal, internal-external and external performance assessment [20, 27–36]. Methodological studies should also make explicit how analytical choices affect reported results and provide reproducible evidence for those choices [37–41]. The unresolved practical question is how these safeguards should be sequenced when rare events, cross-sectional temporality, complex sampling, missingness and cross-setting heterogeneity occur together.

RARE-CROSS addresses the gap between these separate recommendations by turning recurrent failure modes into a single operational sequence for cross-sectional rare-event analyses. It is organised around three verbs: DEFINE the target and estimand, PROTECT the analytical pipeline from avoidable bias and

leakage, and PROVE performance beyond development data. Here, PROVE means generating transparent evidence that performance persists in data excluded from model development, not claiming universal validity. We aimed to develop a framework that can be used prospectively to plan an analysis, retrospectively to audit an existing workflow, and reproducibly to document unresolved cautions. We evaluated key framework claims through targeted simulations and a multi-country WHO STEPS demonstration.

Notes: The nine domains are organised around DEFINE the prediction target, PROTECT the analytical pipeline from leakage and bias, and PROVE performance beyond development data.

Methods

Development, methodological basis and intended use of RARE-CROSS

RARE-CROSS was developed as an evidence-informed methodological framework rather than a consensus reporting guideline. Development proceeded in four stages: (1) identify recurrent failure modes in rare-event prediction, cross-sectional survey analysis, missing-data handling, calibration and transportability; (2) map each failure mode to established methodological guidance; (3) translate the guidance into an operational guardrail with a minimum reporting item and, where useful, an enhanced option; and (4) evaluate claims that could be stress-tested quantitatively using simulations or the empirical STEPS example. A candidate domain was retained when the failure mode could materially alter the estimand, event probability or reported performance; recurred across methodological literature or applied survey workflows; admitted a concrete preventive or audit action; and was not fully subsumed by another domain. These criteria produced nine domains. This approach follows principles for methodological and simulation studies [37–39]. No Delphi process or formal consensus exercise was undertaken, so RARE-CROSS should be interpreted as an operational synthesis whose value depends on transparent application and continued empirical evaluation.

The framework also draws on the authors' applied experience with large health-survey datasets: survey-weighted analysis of cardiovascular disease across African WHO STEPS surveys [42]; an open-source, design-aware R/Python workflow for predictor selection in the Global Youth Tobacco Survey [43]; a practical R tutorial on logistic regression for clinical and public-health research [9]; and a metadata-preserving R workflow for labelled Stata, SPSS and SAS health-survey files [44]. These experiences informed the emphasis on estimand clarity, preservation of data meaning, survey design, leakage-safe preprocessing, calibration, cross-setting heterogeneity and reproducible implementation. RARE-CROSS complements rather than replaces TRIPOD + AI, PROBAST + AI and PRICSSA.

The nine domains are: R1, Research question and estimand; A, Ascertainment, outcome definition and temporal plausibility; R2, Representativeness and complex survey design; E, Event rarity and class-imbalance handling; C, Covariate temporality, leakage and detection bias; R3, Robust missing-data

strategy; O, Out-of-sample geographic transportability; S1, Scoring, calibration and performance evaluation; and S2, Scope of interpretation, transparency and reporting. Table 1 summarizes the guardrail associated with each domain.

Table 1
RARE-CROSS domains and analytical guardrails.

Domain	Methodological focus	Failure mode	RARE-CROSS guardrail	Minimum reporting item
R1	Research question and estimand	Calling prevalent classification future risk prediction	Define task, time origin and intended use	State exactly what probability the model estimates
A	Ascertainment and temporal plausibility	Wrong event orientation or inconsistent outcome	Audit coding, wording, recall period and event class	Report positive class and prevalence
R2	Representativeness and survey design	Ignoring weights, clustering or country composition	Separate population estimation from predictive fitting; run weighting sensitivity	State which design features are incorporated
E	Event rarity and imbalance	Majority-class dominance or synthetic-data leakage	Report event count and prevalence; assess information relative to model complexity; keep any balancing inside training data	Report events, prevalence, imbalance strategy, natural-prevalence evaluation and uncertainty
C	Covariate temporality and detection bias	Post-outcome or healthcare-contact information appears predictive	Preclassify predictor temporality; run the RARE-CROSS detection-dependence stress test for clinically ascertained variables	Do not interpret predictive importance causally
R3	Robust missing-data strategy	Structural absence treated as ordinary missingness	Separate item, structural, derived and informative missingness; impute only inside training resamples	Report country-specific availability and retention
O	Out-of-sample transportability	Random split mistaken for transportability evidence	When transport across settings is intended, hold countries, sites or clusters out of model development	Report performance by held-out setting
S1	Scoring and calibration	AUC-only evaluation or prevalence-blind scoring	Report PR AUC, prevalence, Brier/BSS, calibration and threshold metrics	Benchmark PR and Brier against no-skill baselines; report interval estimates

Domain	Methodological focus	Failure mode	RARE-CROSS guardrail	Minimum reporting item
			with uncertainty; add decision utility when action-linked	and decision utility if relevant
S2	Scope and reporting	Causal or clinical claims unsupported by design	Separate model performance, behaviour, association and utility	Align claims with design, intended use and transportability evidence; provide reproducible audit output

Relationship to existing standards

RARE-CROSS is intended to operationalise, not replace, existing reporting and appraisal standards. The crosswalk below makes the division of labour explicit and clarifies the additional workflow contribution of RARE-CROSS.

Table 2
Relationship between RARE-CROSS and established methodological standards.

Standard / guidance	Primary purpose	Overlap with RARE-CROSS	Distinct RARE-CROSS contribution
TRIPOD + AI / TRIPOD [27, 32]	Transparent reporting of prediction-model studies.	Target population, outcome, predictors, model development/validation and performance reporting.	Orders rare-event survey decisions prospectively; adds explicit guardrails for base-rate interpretation, leakage-safe balancing, survey structure and detection-sensitive predictors.
PROBAST + AI / PROBAST [28, 31]	Risk-of-bias and applicability appraisal.	Participants, predictors, outcomes, analysis, overfitting and applicability.	Turns recurrent bias concerns into prospective actions and an auditable Addressed/Caution/Not-applicable workflow rather than an appraisal score.
PRICSSA [20]	Reporting of complex-sample survey analyses.	Weights, clustering, stratification, target population and design transparency.	Connects survey design to predictive fitting, resampling, evaluation and transportability; distinguishes design-based estimation from weight-sensitivity analyses.
Simulation-study guidance [38, 39]	Design and reporting of simulation studies.	Data-generating mechanisms, estimands, replications, methods and performance measures.	Uses targeted failure-mode simulations to test specific guardrails and exposes when workflow errors manufacture apparent performance.
Validation and dataset-shift guidance [29, 30, 33–36]	Assessment of performance beyond model development data.	Internal-external/external validation, calibration, heterogeneity and dataset shift.	Makes setting-held-out performance a dedicated domain linked to survey heterogeneity, earlier preprocessing decisions and scope of interpretation.

Prior applied use of the framework

Core RARE-CROSS principles were applied in an earlier 12-country WHO STEPS cardiovascular-disease modelling study reported as a medRxiv preprint. That analysis included 60,294 adults and compared elastic-net logistic regression, random forest and XGBoost under a reproducible multicountry pipeline [45]. It provides an applied use case showing that the framework can structure real survey modelling, but it is not independent validation because of author overlap and its preprint status. Experience from that application motivated the present study's stronger stress tests of leakage, calibration, detection-sensitive predictors and cross-setting transportability.

Empirical WHO STEPS demonstration

We used the harmonised analytic file from the accompanying WHO STEPS cardiovascular-disease benchmarking programme. WHO STEPS provides standardised surveillance of behavioural and biological noncommunicable-disease risk factors together with survey-design variables [46]. The broader STEPS resource has previously been used by the authors for survey-weighted estimation of cardiovascular-disease prevalence, prevention and treatment [42] and for the 12-country applied RARE-CROSS use case described above [45]. For the present methodological demonstration, we applied a stricter analysis-specific eligibility rule: African survey years 2014–2019, non-missing harmonised CVD outcome, required common predictors, and both outcome classes within each eligible country. The resulting cohort comprised 52,572 adults from 11 countries, including 2,778 CVD-positive participants (5.28%). This explicit re-derivation is intentional because reproducible prediction work should make the transition from source data to model-ready data auditable.

The outcome was self-reported history of heart attack, angina or chest pain from heart disease, or stroke. Because predictors were measured cross-sectionally, the estimand was the conditional probability of belonging to the prevalent self-reported CVD-positive class at the survey visit. No prospective event horizon was implied. The empirical demonstration was designed to test methodological behaviour and geographic transportability, not to produce a deployable clinical risk model or treatment-decision rule.

The full predictor set comprised age, sex, residence, education, smoking history, harmful alcohol use, physical activity, high salt intake, body mass index, hypertension status, diabetes status and DASH dietary score. The detection-adjusted set removed hypertension and diabetes because those variables may capture both underlying cardiometabolic biology and prior clinical detection. Their removal was treated as a sensitivity analysis, not as proof that all lost signal represented detection bias.

RARE-CROSS implementation workflow

The central pipeline rule is simple: any operation that learns from data belongs inside the training resample. This includes imputation, category encoding, scaling, feature selection, SMOTE, hyperparameter tuning, probability recalibration and threshold selection. The final evaluation fold, external country or other holdout remains untouched until prediction. Figure 2 translates this rule into an implementation sequence that can be added to existing regression or machine-learning pipelines without requiring a new modelling algorithm.

Notes: Operations that learn from the data are nested within training resamples; the final evaluation data remain untouched until prediction. The workflow is algorithm-agnostic and can wrap around statistical or machine-learning classifiers.

How analysts can adopt RARE-CROSS in practice

RARE-CROSS is designed to strengthen an existing workflow rather than replace it. At protocol stage, analysts specify the target, positive event and survey population (R1, A and R2). Before model fitting, they audit event rarity, predictor temporality and detection sensitivity, and the structure of missing data (E, C and R3). During development, all learned preprocessing stays inside training resamples. During evaluation, ROC AUC is paired with PR AUC, calibration and base-rate benchmarks, and setting- or cluster-held-out assessment is used when transport across contexts is part of the intended use (O and S1). Finally, claims are aligned with the design and decision context (S2). This modular structure lowers the barrier to adoption because teams can retain their preferred algorithms and software while improving the credibility of the surrounding analytical process.

Core and enhanced implementation levels

To make the framework usable across teams with different data resources, RARE-CROSS separates a core implementation from enhanced analyses (Table 3). Core items are expected whenever the corresponding domain is relevant; enhanced items deepen stress testing when sample size, survey structure and intended use justify them. This distinction prevents the framework from becoming an all-or-none checklist while preserving the safeguards that protect rare-event analyses.

Table 3
Core and enhanced RARE-CROSS implementation levels.

Domain	Core implementation	Enhanced option when feasible
R1: Question and estimand	Define target population, positive event, time origin/visit and intended use.	Predefine deployment scenario, subgroup estimands and decision thresholds.
A: Ascertainment and temporality	Audit coding, event orientation, wording, recall period and temporal plausibility.	Repeat analyses under credible alternative event definitions or label assumptions.
R2: Representativeness	Describe weights, strata, clusters, country composition and the role of survey design in fitting and evaluation.	Use survey-aware or cluster-respecting resampling and compare weighted with unweighted analyses.
E: Event rarity and imbalance	Report absolute event count and prevalence; keep balancing inside training data; evaluate at natural prevalence.	Compare class weighting and training-only resampling; correct/recalibrate probabilities if class priors were altered.
C: Covariate temporality and detection	Classify predictors by temporal and detection plausibility; avoid causal interpretation of predictive importance.	Run the detection-dependence stress test and add healthcare-contact or ascertainment sensitivity analyses.
R3: Missing data	Separate item, structural, derived and informative missingness; fit imputation only within training resamples.	Use multiple imputation or scenario-based sensitivity analyses that preserve between-setting structure.
O: Out-of-sample transportability	When performance is intended to travel, evaluate settings or clusters excluded from model development.	Use repeated internal-external cross-validation, temporal holdouts or prespecified setting-specific updating.
S1: Scoring and calibration	Report ROC AUC with PR AUC, prevalence, Brier/BSS, calibration and threshold metrics plus uncertainty.	Add calibration updating experiments and decision-curve/net-benefit analysis when predictions will trigger action.
S2: Scope and transparency	Separate predictive performance from causal or clinical claims; disclose preprocessing, seeds, code and intended use.	Produce a one-page audit summary/model card and versioned reproducibility package.

For audit use, each domain can be marked Addressed, Caution, or Not applicable, with one line of supporting evidence and one action if a caution remains. A compact audit summary is therefore informative rather than punitive: it makes unresolved assumptions visible without converting RARE-CROSS into a numerical score.

Prevalence-aware performance metrics

For event prevalence p, sensitivity Se and specificity Sp, positive predictive value can be reconstructed directly from the base rate and operating characteristics:

$$PPV = (p \times Se) / [p \times Se + (1 - p) \times (1 - Sp)]$$

1

The F1 score provides a harmonic summary of precision and sensitivity:

$$F1 = 2 \times PPV \times Se / (PPV + Se)$$

2

For a prevalence-only model predicting p for every participant, the null Brier score is:

$$Brier_{null} = p(1 - p)$$

3

We therefore express probability accuracy relative to the null model using the Brier skill score:

$$BSS = 1 - Brier_{model} / Brier_{null}$$

4

Because the no-skill PR AUC is approximately the positive-event prevalence, we also use PR lift:

$$PR\,Lift = AUC_{PR} / p$$

5

These quantities make comparisons across populations with different prevalence more interpretable than raw PR AUC or Brier score alone.

Event-count adequacy, uncertainty and decision usefulness

Rare-event difficulty is determined by both prevalence and the absolute number of events. A 1% outcome in 500 participants is methodologically different from a 1% outcome in 100 000 participants. RARE-CROSS therefore requires reporting both event prevalence and event count, and asks whether the number of events is adequate for the intended predictor complexity, resampling strategy and performance assessment. It deliberately avoids simplistic events-per-variable rules and instead points investigators to contemporary sample-size and penalisation principles, together with transparent logistic-regression specification and diagnostics [5, 6, 9, 47].

Performance estimates should be accompanied by uncertainty wherever feasible. Bootstrap or other resampling approaches that respect the relevant grouping structure can be used for intervals around discrimination, calibration and threshold-dependent measures; paired methods are preferable when comparing correlated ROC AUCs [14, 48, 49]. For decision-support applications, decision-curve analysis can be added as an enhanced RARE-CROSS option, but it is not mandatory for surveillance or descriptive classification tasks [17, 18].

Simulation experiments

We used targeted Monte Carlo experiments to make specific failure modes visible rather than constructing an exhaustive factorial simulation. Simulation design followed the logic of defining aims, data-generating mechanisms, estimands, methods and performance measures, with fixed replication schedules and seeds for reproducibility [38, 39]. Each experiment generated binary outcomes from a known logistic mechanism with linear, nonlinear and interaction terms. The prevalence experiment used 80 replications per target prevalence of 1%, 2.5%, 5% and 10%; the SMOTE experiment used 60 replications at approximately 2.5%; the detection-dependence experiment used 80 replications at approximately 5%; and the missingness experiment used 50 replications. For each replicated metric we report the mean and, in Supplementary Table S4, the between-replication standard deviation, Monte Carlo standard error (SD/sqrt[R]) and an approximate 95% Monte Carlo interval for the estimated mean (mean +/- 1.96 x MCSE). These intervals quantify simulation precision rather than population sampling uncertainty.

For participant i, the latent event probability was generated from:

$$logit\left[P(Y_i = 1)\right] = \alpha + \beta \square X\square + \beta \square X\square + \beta \square X\square + f(X\square) + \beta \square X\square X\square$$

6

The intercept alpha was numerically tuned to the target prevalence. A clinically ascertained covariate D was generated using:

$$logit\left[P(D_i = 1)\right] = \gamma \square + \gamma \square Y_i + \gamma \square X\square\square$$

7

where gamma1 controlled the strength of outcome-dependent detection. The difference between full and detection-adjusted PR AUC was summarized descriptively as:

$$Detection\ sensitivity = (AUC_PR, full - AUC_PR, adjusted) / AUC_PR, full$$

8

This statistic was interpreted only as sensitivity of model performance to detection-related information, not as the fraction of performance causally attributable to detection bias.

We refer to this paired analysis as the RARE-CROSS Detection-Dependence Stress Test (DDST). The DDST has three steps: identify variables that may encode prior diagnosis or healthcare contact; compare full and detection-adjusted models under exactly the same resampling and evaluation scheme; and report changes in discrimination, calibration and threshold metrics as dependence on detection-related information. The DDST is deliberately noncausal: a performance drop does not quantify the proportion of signal caused by detection bias.

Class imbalance and probability correction

We compared original-prevalence fitting, class weighting and SMOTE applied only to training data. One simulation intentionally applied SMOTE before the train-test split to demonstrate leakage. This deliberately incorrect analysis was confined to simulation. The experiment reflects a broader literature showing that resampling can be useful but is highly sensitive to where it occurs in the analytical pipeline [2–4, 8]. Raw probabilities from the approximately balanced SMOTE training distribution were corrected back to the target prevalence using an odds-based prior correction:

$$logit(p_target) \;=\; logit(p_raw) \;+\; logit(\pi_target) \;-\; logit(\pi_training)$$

9

where pi_training is the event prevalence represented by the model-fitting distribution and pi_target is the prevalence in the target population. Ranking metrics are invariant to this monotonic correction, but absolute probability calibration is not.

A practical decision rule for imbalance handling

1. Start with an unbalanced reference analysis so that the natural event prevalence and baseline probability scale are visible.
2. If minority-class learning is inadequate, compare class weighting and/or resampling performed inside the training data only; never balance the final evaluation set.
3. Evaluate every strategy on untouched data at the target prevalence using PR AUC, PR lift, calibration, Brier/Brier skill and threshold-dependent metrics, not ROC AUC alone.
4. If training altered class priors, treat raw probabilities as potentially distorted and assess prior correction or recalibration using development data only.
5. Select the strategy that improves the performance domain needed for the intended use without sacrificing probability reliability or transportability.

Missing-data experiment

Missingness was separated into participant-level item nonresponse, structural country or module absence, derived-variable missingness and informative measurement availability. This distinction is central because complete-case analysis and imputation can target different populations, and imputation must be integrated with resampling to avoid information leakage [21–24]. In simulation, missingness

was introduced into correlated predictors with probability dependent on observed covariates and the outcome. We compared complete-case analysis, iterative multivariable imputation and marginal stochastic replacement.

RARE-CROSS missingness taxonomy

The framework treats four missingness types differently because they imply different analytical actions:

- **Item nonresponse.** A variable was intended to be measured for the participant but is absent. Imputation can be considered, but the imputation model must be learned inside the training resample.
- **Structural absence.** A variable was never collected in a country, module or wave. This is not ordinary participant missingness; analysts should consider a common-predictor model, setting-specific model, or explicit structural-missingness sensitivity analysis rather than inventing a full setting through imputation.
- **Derived-variable missingness.** The derived feature is unavailable because one or more source variables are missing. Derivation rules and any learned imputation must be reproduced inside resampling to prevent leakage.
- **Informative measurement availability.** Whether a variable was measured may itself depend on healthcare contact, symptoms or survey design. Availability indicators and scenario analyses can be useful, but should be interpreted as part of the data-generating process rather than as neutral missingness.

Geographic transportability assessment and survey-weight sensitivity

The empirical reference model was logistic regression because its probability scale and coefficients make methodological contrasts transparent [9]. Internal performance used a stratified 80:20 participant split. Geographic transportability was then assessed by leave-one-country-out evaluation, so every test country was absent from imputation, encoding and model fitting. This structure is aligned with guidance on internal-external validation, structured cross-validation and dataset shift [29, 30, 33–36]. Country-specific PR AUC was interpreted alongside prevalence and summarised as PR lift.

$$Transportability\ gap\,(M)\ =\ M_internal\ -\ M_external$$

10

Survey-weight sensitivity compared unweighted fitting with fitting using weights normalised to a mean of approximately one within country. Weighted performance was evaluated using the normalised survey weights. This does not convert a machine-learning model into a fully design-based survey estimator; it provides a sensitivity analysis for dependence on unequal sampling weights. The distinction is practically important: previous STEPS analyses by the authors used survey-weighted population

estimation [42], while a GYTS methods study showed that design-based R analyses and weight-only Python implementations can agree closely on point estimates yet differ in uncertainty when clustering and stratification are not represented identically [43]. RARE-CROSS therefore requires analysts to state explicitly which design elements enter estimation, fitting, resampling and evaluation.

Software, reproducibility and adoption resources

The numerical demonstrations were implemented in Python using pandas, NumPy, scikit-learn, imbalanced-learn and statsmodels, with fixed seeds, executable scripts, machine-readable outputs and verification checks. The accompanying R materials provide a parallel implementation using established predictive-modelling and survey-analysis tooling. Reproducibility starts before model fitting: preserving variable labels, value labels and dataset semantics during import reduces recoding error and strengthens the audit trail [44], while open-source cross-platform workflows can make complex-survey analysis more transparent and comparable [43]. The submission includes the implementation workbook and audit checklist (Additional File 1), annotated R code (Additional File 2), Python reproducibility script (Additional File 3), synthetic demonstration data (Additional File 4), high-resolution figure bundle (Additional File 5), supplementary tables and figures including Monte Carlo uncertainty summaries (Additional File 6), and the complete replication bundle with run scripts, requirements, expected results, verification utilities, example outputs, session information and checksums (Additional File 7). These resources link manuscript claims to executable code while keeping restricted WHO participant data outside the public bundle [40, 41].

Reporting standards

Reporting was informed by TRIPOD + AI for prediction-model components, PRICSSA for complex-survey elements and established simulation-study principles; PROBAST + AI informed consideration of bias and applicability [20, 27, 28, 38, 39]. These standards are complementary to RARE-CROSS and remain the appropriate reference standards for reporting and appraisal.

Results

RARE-CROSS converts a chain of validity threats into guardrails

The framework separates the analysis into three linked questions. DEFINE asks whether the target and event are conceptually coherent. PROTECT asks whether preprocessing, imbalance handling, missing-data management and predictor selection preserve the integrity of the evaluation. PROVE asks whether calibrated performance persists when the model leaves the development environment. This organization prevents class imbalance from dominating the methodological conversation when more fundamental errors may already have occurred upstream.

Rare-event metrics changed sharply with prevalence even when ROC AUC did not

Across target prevalence values from 1% to 10%, mean ROC AUC remained within a comparatively narrow range (0.793 at 1% and 0.773 at 10%), whereas PR AUC increased from 0.105 to 0.328 and PPV and F1 also rose with the base rate. At 1% prevalence, the approximate 95% Monte Carlo intervals for the mean were 0.780–0.806 for ROC AUC and 0.091–0.119 for PR AUC; at 10% they were 0.768–0.779 and 0.320–0.337, respectively. PR lift declined from 10.56 to 3.30, illustrating why raw PR AUC should be interpreted against prevalence (Supplementary Table S1; Supplementary Table S4; Supplementary Fig. S1).

Pre-split SMOTE manufactured performance and raw SMOTE probabilities were miscalibrated

At approximately 2.5% prevalence, SMOTE applied correctly inside the training data produced mean PR AUC 0.166 (95% Monte Carlo interval for the mean 0.151–0.180) and ROC AUC 0.792. Applying SMOTE before the split contaminated the evaluation distribution and produced an apparent PR AUC of 0.799 (0.793–0.805) and ROC AUC 0.816. The magnitude of the PR inflation demonstrates why resampling placement is a validity issue rather than a minor coding detail (Supplementary Table S4; Supplementary Fig. S2).

Raw SMOTE probabilities were also severely miscalibrated on the original-prevalence test data: mean Brier score was 0.178 (95% Monte Carlo interval for the mean 0.175–0.180) and the calibration intercept was − 3.645. Prior-prevalence correction reduced the mean Brier score to 0.023 (0.022–0.023) and shifted the calibration intercept to -0.450 without changing PR AUC or ROC AUC. Ranking and probability reliability therefore need separate evaluation.

RARE-CROSS detection-dependence stress test identified clinically ascertained signal

In the controlled detection-dependence simulation, adding a covariate whose ascertainment depended partly on the outcome increased mean ROC AUC from 0.779 to 0.832 and mean PR AUC from 0.224 (95% Monte Carlo interval for the mean 0.216–0.231) to 0.327 (0.317–0.336). The mean DDST statistic was 0.315. This quantifies sensitivity to detection-related information under the specified data-generating mechanism; it is not a causal attribution (Supplementary Table S4).

The empirical CVD demonstration showed the same direction. In the random-split unbalanced reference model, the full predictor set achieved ROC AUC 0.607 and PR AUC 0.085; removing hypertension and diabetes reduced these to 0.574 and 0.071. Clinically ascertained variables therefore carried meaningful signal, but the framework deliberately leaves open whether that signal reflects biology, diagnosis, treatment, healthcare contact, or a combination (Supplementary Fig. S3).

Missing-data method changed the evaluated population

Under outcome-dependent missingness, complete-case analysis retained on average 893 of 1,200 evaluation observations (95% Monte Carlo interval for the mean 890–897) and reduced event prevalence in the analysed subset to approximately 1.68% (1.53%-1.82%), despite a data-generating prevalence of about 5%. Mean PR AUC was 0.112 for complete-case analysis, 0.111 for iterative multivariable imputation and 0.122 for marginal replacement; the corresponding 95% Monte Carlo intervals for the means were 0.097–0.127, 0.105–0.118 and 0.114–0.130. The principal lesson is therefore not that one missing-data method universally dominates, but that complete-case restriction can change the evaluated population and event base rate before model performance is calculated (Supplementary Table S4; Supplementary Fig. S4).

The African STEPS case study exposed heterogeneity hidden by a single pooled estimate

Country prevalence ranged from 3.3% in Morocco to 9.3% in Uganda. In leave-one-country-out evaluation, country-specific PR AUC ranged from 0.050 in Zambia to 0.163 in Malawi. Pooled out-of-country predictions yielded ROC AUC 0.584, PR AUC 0.079, Brier score 0.050 and calibration slope 0.742. Country-specific calibration intercepts ranged from − 1.532 to 1.583 and slopes from 0.408 to 1.402, showing heterogeneity concealed by pooled performance. Country cohort details are in Supplementary Table S2 and PR-lift results in Supplementary Fig. S5.

Class weighting changed the probability scale more than discrimination

In the empirical random-split logistic model, class weighting changed ROC AUC only from 0.607 to 0.608 and left PR AUC near 0.085, but the raw Brier score worsened from 0.050 to 0.238 and the calibration intercept shifted from − 0.302 to -2.887. Class weighting can therefore change optimisation emphasis without preserving population probability calibration. Selected empirical contrasts are reported in Supplementary Table S3.

Survey-weight sensitivity was modest for discrimination but not conceptually optional

Using normalised survey weights for fitting changed weighted ROC AUC from 0.6127 to 0.6134, weighted PR AUC from 0.0802 to 0.0820 and weighted F1 from 0.149 to 0.154. The numerical changes were modest in this demonstration, but the conceptual requirement remains: analysts should state how design features enter description, fitting, resampling and evaluation because weight sensitivity is data dependent [19, 20, 25, 26].

Discussion

RARE-CROSS starts from a simple proposition: rare-event modelling from cross-sectional surveys is not merely a class-imbalance problem. It is simultaneously a problem of estimand definition, ascertainment, temporality, representativeness, missingness, calibration and transportability. Its methodological contribution is to place these validity threats in an order that analysts can act on and audit. The architecture reflects recurrent challenges encountered in survey-weighted STEPS analyses, complex-survey workflows, logistic-regression practice and reproducible survey-data processing [9, 42–44]. The framework has also been used in a related 12-country WHO STEPS CVD modelling preprint [45], providing an applied feasibility example. Because that study shares authors and remains a preprint, it should not be interpreted as independent validation; rather, it shows how the guardrails can structure a real multicountry pipeline and helped motivate the more explicit stress tests reported here.

The first lesson is that prevalence belongs inside performance interpretation. ROC AUC can look stable while precision changes sharply, so rare-event analyses should report PR AUC together with evaluation prevalence and, where useful, prevalence-standardised PR lift [10, 11]. Brier scores should likewise be interpreted against a prevalence-only benchmark, and calibration should be reported on its own terms rather than inferred from discrimination [12–16].

The second lesson is that balancing is an optimisation choice, not a validity repair. SMOTE, weighting and other imbalance methods can improve minority-class learning, but they cannot repair a misdefined outcome, structural selection, temporal ambiguity or leakage [1–4, 7, 8]. Our pre-split SMOTE experiment produced extreme apparent PR performance because synthetic information entered the evaluation distribution. Even correctly nested balancing can alter class priors and therefore the probability scale, which is why calibration must be assessed separately from ranking.

The third lesson concerns covariate temporality. In cross-sectional disease classification, a clinically ascertained predictor may be predictive because it is biologically related to disease, because the disease prompted diagnosis, because healthcare contact generated both labels, or because these mechanisms coexist. Feature importance alone cannot determine why that signal exists. The RARE-CROSS Detection-Dependence Stress Test therefore compares a full model with a detection-adjusted model under identical resampling. In both simulation and the STEPS demonstration, performance fell after removing detection-sensitive variables. That decline demonstrates dependence on information carried by those variables; it does not show that the difference was caused entirely by detection bias.

The fourth lesson is that missingness is part of the estimand and the data-generating process. Complete-case analysis can change both sample composition and event prevalence; structural non-collection across countries is even more problematic because imputing an entirely unmeasured variable can create pseudo-information. RARE-CROSS therefore requires a variable-availability audit before imputation and keeps learned imputation inside development resamples [21–24].

The fifth lesson is that complex survey design cannot be reduced to whether a modelling function accepts a weight argument. Survey weights address unequal selection and calibration to target populations, while clustering and stratification affect the dependence structure and uncertainty. This

distinction has been operationalised in the authors' survey-weighted STEPS work [42] and in an open-source GYTS methodology that explicitly compared design-based R inference with Python implementations [43]. RARE-CROSS therefore separates design-based population description from predictive fitting and makes weighting sensitivity explicit when fully survey-aware learners are unavailable [19, 20, 25, 26, 42, 43].

The sixth lesson is that internal performance does not establish transportability. Random participant splitting allows development and evaluation data to share countries, survey procedures and healthcare context. Country-held-out evaluation is a stronger stress test for intended use across settings, and heterogeneity should be interpreted as evidence about dataset shift rather than as a fixed property of the algorithm [29, 30, 33–36]. The STEPS results illustrate this directly: pooled discrimination concealed substantial between-country differences in PR AUC and calibration. RARE-CROSS therefore makes setting-held-out performance and calibration central whenever users intend a model to travel.

RARE-CROSS is algorithm-agnostic. Logistic regression, penalised regression, random forests, gradient boosting and other supervised learners can sit inside the same guardrails; the framework governs how data are defined, partitioned, preprocessed, balanced, evaluated and reported rather than prescribing a particular learner [50]. This makes it suitable both for transparent statistical baselines and for more complex machine-learning pipelines.

The framework is intentionally compatible with established reporting and appraisal standards. TRIPOD + AI and TRIPOD guide reporting; PROBAST + AI and PROBAST support bias and applicability assessment; PRICSSA focuses on complex surveys; and external-validation guidance addresses performance beyond the development sample [20, 27–32]. RARE-CROSS adds the operational sequencing needed when these issues occur together in a low-prevalence cross-sectional analysis.

What RARE-CROSS is and is not

RARE-CROSS is an analytical design and audit architecture for rare-event models built from cross-sectional survey data. It is not a new classifier, a causal-inference framework, a replacement for TRIPOD + AI or PROBAST + AI, a numerical quality score, or a guarantee that a fitted model will transport successfully. Its purpose is to order the decisions that determine whether reported performance is credible and to make those decisions auditable. This distinction allows immediate use without implying that the framework itself is a prediction model requiring separate external validation.

Practical adoption and implementation

An implementation advantage of RARE-CROSS is that teams can retain their preferred algorithms while strengthening the surrounding analytical process. The minimum useful implementation is explicit: verify event coding and estimand; preserve data semantics from import; audit structural versus participant-level missingness; keep learned preprocessing inside resampling; report prevalence-aware

discrimination and calibration; stress-test detection-sensitive predictors and survey weights where relevant; and evaluate the settings to which performance is intended to travel. Prior open-source workflows demonstrate that reproducible label preservation and cross-platform survey modelling can be implemented with tools accessible to public-health analysts [43, 44]. The supplied checklist, code, synthetic data and automated verification resources make the same principles directly reusable in RARE-CROSS applications.

A practical implementation can finish with a one-page RARE-CROSS audit summary. For each of the nine domains, analysts record the status (Addressed, Caution, or Not applicable), the evidence supporting that status, and any action required before deployment or interpretation. For example, an analysis may be strong on event definition, leakage control and calibration but retain a caution for structural missingness or geographic transportability. This output is intended to function like a methodological model card for analysts, reviewers and adopters rather than a pass/fail score.

Strengths and limitations

Strengths of this study include explicit mapping of operational guardrails to established methodological standards, transparent criteria for retaining the nine domains, a direct crosswalk showing where RARE-CROSS complements rather than duplicates TRIPOD + AI, PROBAST + AI and PRICSSA, simulations with known data-generating mechanisms, Monte Carlo precision summaries, a multi-country empirical demonstration, an earlier applied use case, separate assessment of discrimination and calibration, and executable R/Python materials that allow numerical claims to be checked. Several limitations should also be emphasised. First, RARE-CROSS is an applied methodological framework, not a consensus reporting guideline or psychometric instrument. Second, the simulations are targeted stress tests rather than an exhaustive factorial comparison of imbalance, imputation or weighting methods [38, 39]; the reported Monte Carlo intervals quantify precision of the simulated means but do not substitute for broader scenario coverage. Third, the empirical example concerns prevalent self-reported CVD, not prospective clinical risk prediction. Fourth, logistic regression was used for transparent methodological contrasts; the framework is algorithm-agnostic but the manuscript does not benchmark every learner. Fifth, the prior 12-country application has author overlap and preprint status and therefore is not independent validation. Sixth, study planning still requires context-specific sample-size calculations; rare events should not be reduced to a fixed events-per-variable rule [5, 6, 47]. Finally, the empirical illustration used a restricted set of countries and years and should not be interpreted as a clinical prediction model for deployment.

Future work should evaluate the framework across broader data-generating conditions and applications, including different event counts, signal strengths, label error, concept shift, structural missingness and repeated internal-external validation. End-to-end nonlinear demonstrations and applications to maternal complications, rare infectious outcomes, injury surveillance, severe mental-health outcomes and disability would further test the generality of the guardrails. Such extensions would expand the evidence

base without changing the central requirement: analytical decisions that affect credibility should be explicit, testable and reproducible.

Conclusions

Rare-event modelling in cross-sectional population surveys requires safeguards beyond ordinary class-imbalance handling. RARE-CROSS converts these safeguards into an actionable sequence organised around DEFINE, PROTECT and PROVE: define the estimand and positive event; protect the analytical pipeline from leakage, detection-sensitive interpretation, structural missingness and probability distortion; and evaluate transportability in data excluded from model development. Targeted simulations showed that prevalence, pre-split oversampling and altered training priors can materially transform apparent performance, while the African WHO STEPS demonstration exposed detection dependence, country heterogeneity and survey-weight sensitivity. RARE-CROSS is designed for immediate, algorithm-agnostic use alongside TRIPOD + AI, PROBAST + AI and PRICSSA. Its contribution is to make the analytical decisions that determine rare-event model credibility visible, auditable and reproducible.

Abbreviations

AUC
area under the curve
BSS
Brier skill score
CVD
cardiovascular disease
DDST
Detection-Dependence Stress Test
MAR
missing at random
ML
machine learning
PPV
positive predictive value
PR AUC
precision-recall area under the curve
PRICSSA
Preferred Reporting Items for Complex Sample Survey Analysis
PROBAST + AI
Prediction model Risk Of Bias ASsessment Tool plus Artificial Intelligence
SMOTE
Synthetic Minority Over-sampling Technique
STEPS

STEPwise approach to noncommunicable disease risk factor surveillance
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence.

Declarations

Ethics approval and consent to participate

This methodological study used harmonised secondary WHO STEPS data supplied under the applicable data-access arrangements. The original national STEPS surveys were conducted under the relevant country-specific ethics approvals and WHO oversight, with informed consent obtained from participants. The present empirical analysis used de-identified records, involved no new participant recruitment or contact, and made no attempt to re-identify participants. Any additional institutional determination required for secondary analysis remained subject to the participating investigators' local requirements.

Consent for publication

Not applicable.

Clinical trial number: not applicable.

Availability of data and materials

The participant-level WHO STEPS data are not redistributed because access is governed by WHO and country-specific data-use conditions. Researchers may request access through the applicable WHO STEPS procedures. All non-restricted replication materials, including the R and Python code, synthetic demonstration data, supplementary outputs, figures, verification files and documentation, are publicly available on Zenodo: https://zenodo.org/records/22846345. Reviewers without STEPS access can reproduce the synthetic-data workflow and simulation checks, while authorised researchers can rerun the empirical pipeline using the documented private-data input structure.

Competing interests

The authors declare no competing interests.

Funding

No specific funding was received for this work.

Authors' contributions

WFN conceived RARE-CROSS; developed the framework and analytical workflow; curated and harmonised the empirical data; designed and executed the simulations and empirical analyses;

developed the reproducibility code and supplementary resources; prepared the visualisations; interpreted the findings; and drafted the original manuscript. CZ contributed to methodological refinement, interpretation of the findings, critical appraisal of the framework and manuscript, and substantive revision of the text. LK contributed senior statistical and methodological input, methodological supervision, interpretation of the results, critical revision of the framework and manuscript, and overall scientific supervision. All authors reviewed and approved the final manuscript and agree to be accountable for the work.

Acknowledgements

The authors acknowledge the national teams, participants and World Health Organization personnel involved in WHO STEPS surveys, and the methodological and reporting-guideline communities whose work informed the framework.

References

1. King G, Zeng L. Logistic regression in rare events data. Political Anal. 2001;9(2):137–63. 10.1093/oxfordjournals.pan.a004868.
2. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863–905. 10.1613/jair.1.11192.
3. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84. 10.1109/TKDE.2008.239.
4. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–32. 10.1007/s13748-016-0094-0.
5. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163. 10.1186/s12874-016-0267-3.
6. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868. 10.1136/bmj.h3868.
7. Zhu J, Pu S, Fang C, et al. Improving rare event prediction in imbalanced clinical data using cost-sensitive logistic regression. BMC Med Res Methodol. 2026. 10.1186/s12874-026-02970-6.
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57. 10.1613/jair.953.
9. Ng'ambi WF, Zyambo C, Muula AS. A practical guide to key considerations in logistic regression for clinical and public health research: R tutorial. Malawi Med J. 2025;37. https://www.mmj.mw/?p=13527. Accessed 19 Sep 2026.
10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.

10.1371/journal.pone.0118432.

11. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006. pp. 233–240. 10.1145/1143844.1143874
12. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
13. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. 10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.
14. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38. 10.1097/EDE.0b013e3181c30fb2.
15. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. 10.1016/j.jclinepi.2015.12.005.
16. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc Mach Learn Res. 2017;70:1321–30.
17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74. 10.1177/0272989X06295361.
18. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. 10.1136/bmj.i6.
19. MacNell N, Feinstein L, Wilkerson J, et al. Implementing machine learning methods with complex survey data: lessons learned on the impacts of accounting sampling weights in gradient boosting. PLoS ONE. 2023;18(1):e0280387. 10.1371/journal.pone.0280387.
20. Seidenberg AB, Moser RP, West BT. Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA). J Surv Stat Methodol. 2023;11(4):743–57. 10.1093/jssam/smac040.
21. Mertens BJA, Banzato E, de Wreede LC. Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: methodological approach and data-based evaluation. Biom J. 2020;62(3):724–41. 10.1002/bimj.201800289.
22. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. 10.1136/bmj.b2393.
23. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920–31. 10.1002/sim.3944.
24. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Boca Raton: CRC; 2018.
25. Lumley T. Complex Surveys: A Guide to Analysis Using R. Hoboken: Wiley; 2010. 10.1002/9780470580066.
26. Heeringa SG, West BT, Berglund PA. Applied Survey Data Analysis. 2nd ed. Boca Raton: CRC; 2017.

27. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
28. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. 10.1136/bmj-2024-082505.
29. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.
30. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7. 10.1016/j.jclinepi.2015.04.005.
31. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. 10.7326/M18-1376.
32. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63. 10.7326/M14-0697.
33. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89. 10.1016/j.jclinepi.2014.06.018.
34. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29. 10.1111/ecog.02881.
35. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit. 2012;45(1):521–30. 10.1016/j.patcog.2011.06.019.
36. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345–52. 10.1093/biostatistics/kxz041.
37. Mbuagbaw L, Lawson DO, Puljak L, Allison DB, Thabane L. A tutorial on methodological studies: the what, when, how and why. BMC Med Res Methodol. 2020;20:226. 10.1186/s12874-020-01107-7.
38. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. 10.1002/sim.8086.
39. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279–92. 10.1002/sim.2673.
40. Peng RD. Reproducible research in computational science. Science. 2011;334(6060):1226–7. 10.1126/science.1213847.
41. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. 10.1371/journal.pcbi.1003285.
42. Ng'ambi WF, Estill J, Merzouki FA, Zyambo C, Banda JC, Beran D, et al. The prevalence, prevention, and treatment of cardiovascular diseases in Twelve African Countries (2014–2019): an analysis of

the World Health Organisation STEPwise approach to chronic disease risk factor surveillance. PLoS ONE. 2026;21(3):e0320276. 10.1371/journal.pone.0320276.

43. Ng'Ambi W, Zyambo C, Kazembe L. Methodological Guidance for Predictor Variable Selection for Adolescent Smoking Outcomes in Global Youth Tobacco Survey Using R and Python. Front Appl Math Stat. 2026;12:1895038. 10.3389/fams.2026.1895038.
44. Ng'ambi WF, Muula AS. A reproducible R workflow to preserve variable and value labels in Stata, SPSS, and SAS datasets for transparent and reproducible health research. Malawi Med J. 2025;37(3):193–206. 10.4314/mmj.v37i3.10.
45. Ng'ambi WF, Merzouki FA, Estill J, Keiser O, Orel E. Machine learning-based prediction of cardiovascular disease risk in Africa using WHO Stepwise Surveys: 2014–2019. medRxiv [Preprint]. 2026:2026.02.23.26346870. 10.64898/2026.02.23.26346870
46. World Health Organization. WHO STEPwise approach to noncommunicable disease risk factor surveillance (STEPS): manual and supporting materials. Geneva: World Health Organization; accessed 2026.
47. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. 10.1136/bmj.m441.
48. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
50. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013. 10.1007/978-1-4614-6849-3.

Figures

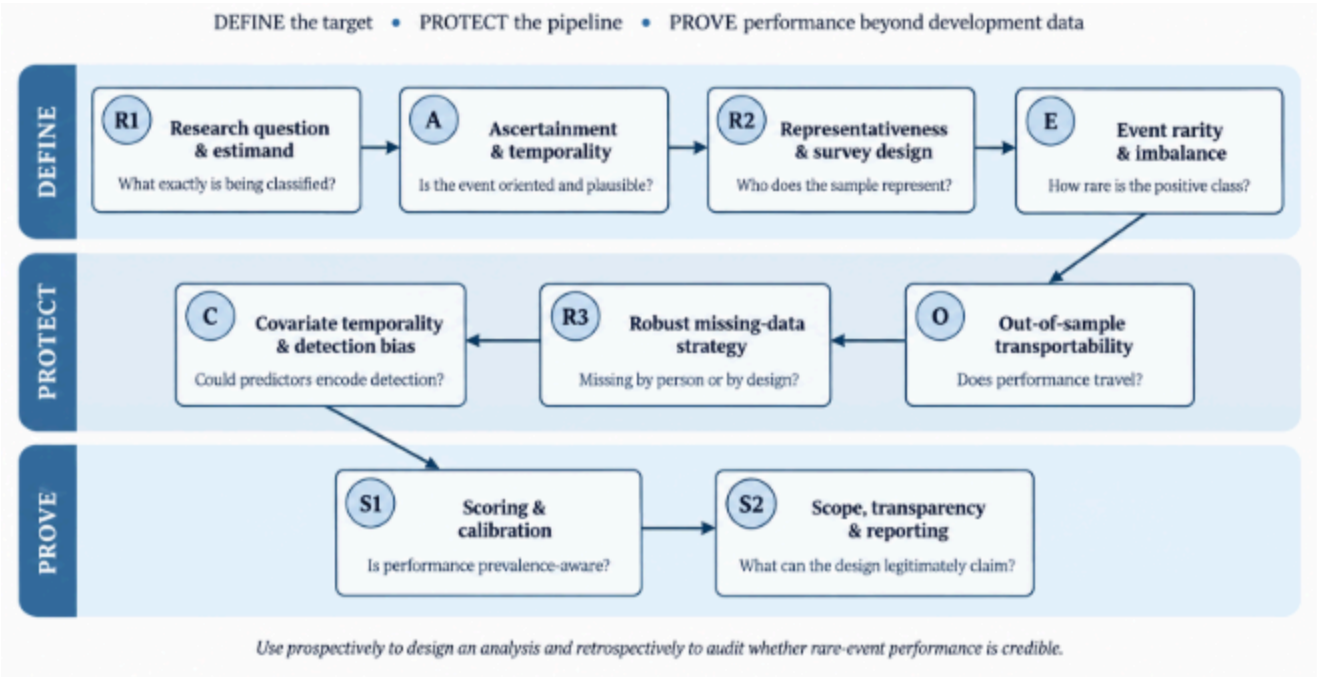


Figure 1

RARE-CROSS framework.

Notes: The nine domains are organised around DEFINE the prediction target, PROTECT the analytical pipeline from leakage and bias, and PROVE performance beyond development data.

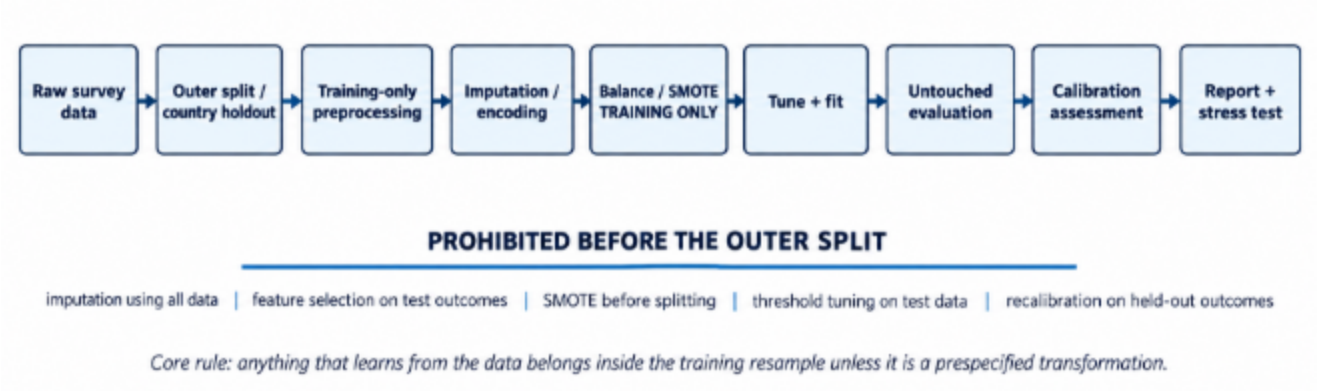


Figure 2

Leakage-safe rare-event modelling workflow.

Notes: Operations that learn from the data are nested within training resamples; the final evaluation data remain untouched until prediction. The workflow is algorithm-agnostic and can wrap around statistical or machine-learning classifiers.

Supplementary Files

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

- RARECROSSSupplementaryTablesandFiguresFINALV6.docx
- RARECROSSSupplementaryWorkbookV6.xlsx