Current Biomedicine
p-ISSN: 2962-8490; e-ISSN: 2985-4784
https://journal.ipb.ac.id/index.php/currbiomed

Adetunji *et al.*
Vol 5 No 1, January 2026, Pp 7–23
DOI: 10.29244/currbiomed.5.1.7

Research **OPEN ACCESS**

Design-consistent statistical learning for rare survey outcomes: a cardiovascular-kidney-metabolic multimorbidity plasmode benchmark in women

Sunday A. Adetunji [1]*, Rhoda O. Oyewusi [2], Tomiisin Balogun [3], Bukola Adetunji [4], Oluwatimileyin C. Adetunji [5]

[1]Department of Biostatistics and Epidemiology, College of Health, Oregon State University, Corvallis, OR, USA; [2]Department of Nursing Science, College of Medicine, University of Lagos, Lagos, Nigeria; [3]Department of Public Health, University of Suffolk, Ipswich, United Kingdom; [4]Directorate of Medical Services and Disease Control, Lagos State Primary Health Care Board, Lagos, Nigeria; [5]Department of Economics, College of Management and Social Sciences, Bowen University, Iwo, Osun State, Nigeria.

Abstract

Background Cardiovascular-kidney-metabolic (CKM) disease integrates metabolic, renal, and cardiovascular systems. Learning in probability samples is challenged by rare outcomes, unequal sampling, clustered dependence, and miscalibration. This study aimed to develop a design-consistent, calibration-first framework for concurrent CKM multimorbidity classification in complex surveys and quantify the consequences of ignoring survey design.

Methods We analyzed the National Health and Nutrition Examination Survey (NHANES) 2017–March 2020 data for U.S. women aged 20–44 years. CKM multimorbidity was defined as the coexistence of at least two of the following: hypertension, diabetes mellitus, chronic kidney disease, and clinical cardiovascular disease. Five learners were compared using 20×5 whole-PSU design-aware cross-validation, fold-contained preprocessing, survey-weighted loss functions, and cross-fitted sigmoid recalibration. The evaluation used design-based and survey-bootstrap procedures across key performance metrics. A 6,000-replicate survey-plasmode benchmark was used to evaluate bias, RMSE, and 95% coverage against finite-population truth.

Results Among 1,743 classifiable women, 44 unweighted cases yielded a survey-weighted prevalence of 1.97% (95% CI, 1.30%–2.65%). The elastic net minimized the calibrated weighted log loss (0.089) and achieved a ROC AUC of 0.797 (95% CI, 0.765–0.830), PR AUC of 0.062 (0.023–0.100), and calibration slope of 0.90. Naive random cross-validation yielded an absolute log-loss bias of 0.0539 and 30.2% coverage; weighted design-aware cross-validation with recalibration achieved 95.0% coverage.

Conclusion For rare CKM outcomes in complex surveys, valid population-level performance estimation requires survey-aware resampling, proper scoring rules, cross-fitted calibration, and design-consistent uncertainties. The framework distinguishes investigator-defined multimorbidity from guideline-aligned CKM staging.

Keywords cardiovascular-kidney-metabolic syndrome | calibration | complex surveys | multimorbidity | statistical learning.

Introduction

Cardiovascular–kidney–metabolic (CKM) multimorbidity is a systemic clinical disorder defined by pathophysiological interactions and bidirectional crosstalk among excess or dysfunctional adiposity, metabolic risk factors (such as type 2 diabetes and dyslipidemia), chronic kidney disease (CKD), and cardiovascular disease (CVD). CKM multimorbidity syndrome

*Corresponding author Email: adetunjs@oregonstate.edu

©The Author(s) 2027 This article is licensed under a Creative Commons Attribution (CC BY 4.0) International License, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, and indicate if changes were made.

provides a unified clinical framework for the interdependence of metabolic risk, chronic kidney disease (CKD), and cardiovascular disease (CVD). The American Heart Association defined a five-stage construct: stage 0, absence of CKM risk factors; stage 1, excess or dysfunctional adiposity; stage 2, metabolic risk factors, CKD, or both; stage 3, subclinical CVD or an equivalent high-risk state; and stage 4, clinical CVD (Ndumele *et al.*, 2023a; Ndumele *et al.*, 2023b). The 2026 AHA/ACC/ADA/ASN clinical practice guidelines retained this staged model and integrated prevention, detection, risk assessment, and treatment across the cardiovascular, kidney, and metabolic domains (Ndumele *et al.*, 2026).

The burden of population is substantial. In a nationally representative analysis of 10,762 U.S. adults from NHANES 2011–March 2020, only 10.6% met the stage 0 criteria, whereas 25.9% were in stage 1, 49.0% in stage 2, 5.4% in stage 3, and 9.2% in stage 4; thus, almost 90% of U.S. adults had CKM stage 1 or higher, and 14.6% had advanced stage 3 or 4 disease (Aggarwal *et al.*, 2024). Among adults aged 20–44 years, the age range most relevant to the present investigation, only 18.2% were in stage 0, and 2.1% already had advanced CKM stages. These estimates did not improve significantly between 2011 and 2020. Social patterning is also evident: in the NHANES 1999–2018, adults with at least two adverse social determinants of health had an age-standardized prevalence of advanced CKM disease of 15.8%, compared with 10.5% among those with fewer adverse social determinants (Zhu *et al.*, 2024).

Reproductive history provides additional clinical information during a period when overt CKM disease remains uncommon. Pregnancy complications are associated with subsequent cardiovascular, metabolic, renal, and multimorbidity outcomes, supporting their role as markers of long-term maternal health rather than as isolated obstetric events (Okoth *et al.*, 2020; Parikh *et al.*, 2021; O'Kelly *et al.*, 2022; McNestry *et al.*, 2023). In a population-based cohort, women with a history of hypertensive disorders of pregnancy had a higher risk of stroke (HR 2.27, 95% CI 1.37–3.76), coronary artery disease (HR 1.89, 95% CI 1.26–2.82), CKD (HR 2.41, 95% CI 1.54–3.78), and accumulation of multimorbidity (HR 1.25, 95% CI 1.15–1.35) than women with normotensive pregnancies (Garovic *et al.*, 2020). A meta-analysis of 23 studies including 5.77 million women evaluated the long-term risk of renal disease following adverse pregnancy outcomes (Barrett *et al.*, 2020). Based on this large-scale population data, women with a history of preeclampsia have an approximately fivefold increase in the risk of developing end-stage kidney disease. Similarly, gestational hypertension is associated with a more than three-fold increase in risk, and a history of preterm delivery more than doubles the risk of subsequent kidney failure. Gestational diabetes has also been associated with later cardiovascular disease, and an updated umbrella review reported a pooled 21% higher risk of stroke after gestational diabetes (RR 1.21, 95% CI 1.05–1.40) (Okoth *et al.*, 2020). Collectively, these findings provide a clinical rationale for evaluating reproductive history, along with socioeconomic and mental-health characteristics when studying the early CKM burden in women.

The methodological problem is distinct from clinical questions. NHANES is generated through multistage, stratified, clustered probability sampling with unequal selection probabilities; consequently, the observed sample and the finite population represented by the sampling weights are different inferential targets. Conventional machine-learning workflows generally assume independent and identically distributed observations and can therefore be incompatible with design-based population inference when survey weights, strata, and primary sampling units are ignored. Wieczorek *et al.* developed the KK-fold cross-validation specifically for complex sample surveys and showed, through mathematical arguments and simulations, that fold construction and test-error estimation should account for stratification, clustering, and unequal selection probabilities when the target is design-based inference (Wieczorek *et al.*, 2022).

Empirical evidence demonstrates that this distinction can materially change the apparent predictive performance. In an NHANES III gradient-boosting analysis, MacNell *et al.* reported an F1 score of 77.4% under the survey-weighted reference analysis compared with 81.9% when the sampling weights were ignored – a five-percentage-point increase in apparent performance. Re-evaluating the predictions with the sampling weights reduced the F1 score to 77.0%, which was close to the fully weighted analysis (MacNell *et al.*, 2023). More recent NHANES work has extended design-based cross-validation and replicate-weight methods to the elastic net, random forest, and XGBoost, but explicitly cautioned that its class-balancing approach was unsuitable for extreme events such as rare diseases (Kim *et al.*, 2026). This leaves an important methodological gap at the intersection of complex survey sampling and rare-outcome statistical learning.

Rare outcomes require particular attention to the estimation of probabilities. A model can discriminate between cases and non-cases while producing poorly calibrated probabilities; therefore, ROC AUC alone is insufficient for evaluating a model intended to estimate absolute risk or prevalence probabilities. Calibration-in-the-large, calibration slope, proper scoring rules such as log loss and the Brier score, and measures of calibration error quantify the complementary aspects of predictive accuracy (Steyerberg *et al.*, 2010; Van Calster *et al.*, 2016; Van Calster *et al.*, 2019). Van Calster *et al.* emphasized that calibration concerns the agreement between predicted probabilities and observed outcome frequencies, and contemporary prediction guidance recommends resampling-based internal validation in preference to arbitrary development–test splitting, which discards information and can produce unstable performance estimates (Collins *et al.*, 2024a; Efthimiou *et al.*, 2024; Riley *et al.*, 2025). These issues are magnified when events are uncommon because a small number of outcomes must support the model estimation, hyperparameter selection, calibration, subgroup assessment, and uncertainty quantification.

Accordingly, this study addresses both clinical estimation and statistical validation problems. We evaluated concurrent CKM multimorbidity among U.S. women aged 20–44 years, a demographic selected to capture women of reproductive age, where early-onset cardiometabolic risks critically impact long-term and maternal health and operationalized as the coexistence of at least two of hypertension, diabetes mellitus, CKD, and clinical CVD, while reporting the 2026 guideline-

aligned CKM stages separately. The primary methodological objective was to evaluate a design-consistent, calibration-first statistical-learning strategy for rare binary outcomes in complex surveys. The framework combines whole-PSU repeated cross-validation, preprocessing restricted to training folds, survey-weighted model estimation and proper scoring rules, cross-fitted probability recalibration, design-consistent uncertainty estimation, and a survey-plasmode experiment with a known finite-population truth. The plasmode benchmark was used to quantify the bias, root mean squared error, and confidence-interval coverage when sampling weights, cluster structure, and calibration were progressively incorporated. In this formulation, the NHANES provides the empirical CKM application, whereas the principal methodological estimand is the accuracy with which alternative validation strategies recover population-level predictive performance under rare-event complex sampling.

Methods

Ethics

This study used public-use, de-identified NHANES data collected under the National Center for Health Statistics research-ethics review and informed-consent procedures. No participant recruitment, contact, intervention, or re-identification was undertaken for this secondary analysis.

Study design, data source, and estimand

We conducted a cross-sectional, design-based statistical-learning study using the National Health and Nutrition Examination Survey (NHANES) pre-pandemic public-use files from 2017 to March 2020. NHANES uses a multistage, stratified, clustered probability design to represent the U.S. civilian non-institutionalized population. The target population for the empirical application was women aged 20–44 years old. Because exposure and outcome information were observed at the same survey assessment, the empirical estimand was the target-population concurrent probability of CKM multimorbidity, rather than future incidence, prognosis, or a causal effect.

The analysis used 18 official NHANES modules covering demographics, reproductive health, depressive symptoms, medical conditions, blood pressure, diabetes, Anthropometry, smoking, glycated hemoglobin, fasting glucose, triglycerides, HDL cholesterol, total cholesterol, serum biochemistry, urine albumin-creatinine ratio, and prescription medications. All modules were linked by a unique participant identifier (SEQN), and duplicate identifiers were prohibited before merging.

For the primary analysis, the parent complex-survey design was constructed among participants with positive combined-cycle MEC examination weights using WTMECPRP as the sampling weight, SDMVSTRA as the masked variance stratum, and SDMVPSU as the masked primary sampling unit (PSU), consistent with the NCHS analytic guidance for the 2017–March 2020 pre-pandemic files (NCHS, 2022). Women aged 20–44 years were analyzed as a survey domain after the construction of the parent design. For CKM staging, the parent design was reconstructed with the fasting-subsample weight WTSAFPRP, because fasting glucose and triglyceride levels contributed to staging. Single-PSU strata arising within restricted domains were handled using the predefined lonely-PSU adjustment.

CKM multimorbidity outcome definition and ascertainment

The primary epidemiological endpoint was CKM multimorbidity, defined as the coexistence of at least two of four disease domains: clinician-diagnosed hypertension, clinician-diagnosed diabetes mellitus, chronic kidney disease (CKD), and clinical cardiovascular disease (CVD). Hypertension was obtained from BPQ020; diabetes mellitus from DIQ010; CKD from KIQ022; and clinical CVD from the composite of congestive heart failure, coronary heart disease, angina, myocardial infarction, or stroke (MCQ160B–MCQ160F). DIQ010 explicitly distinguishes diabetes diagnosed by a health professional from diabetes occurring only during pregnancy; consequently, gestational-only diabetes was not classified as current diabetes.

The outcome components were coded as present, absent, or unknown. Refused, uncertain, and non-response codes were retained as unknown rather than recoded as disease absence. For participant i, let P_i denote the number of known-positive CKM domains and U_i the number of domains with unknown status. The tri-state endpoint-ascertainment for participant i (Y_i) is defined as:

$$Y_i^{\mathrm{CKM}} = \begin{cases} 1, & P_i \geq 2, \\ 0, & P_i + U_i < 2, \\ \text{unclassified}, & P_i < 2 \leq P_i + U_i. \end{cases} \quad (1)$$

where Y_i^{CKM} represents the final tri-state CKM multimorbidity classification status for participant i, P_i denotes the count of known-positive clinical disease domains for that individual, and U_i represents the count of domains with unknown or missing clinical status.

Thus, a participant was a definite case when at least two domains were known to be positive $P_i \geq 2$, a definite non-case when even assigning every unknown domain as positive could not reach the threshold $P_i + U_i < 2$, and otherwise remained unclassified $P_i < 2 \leq P_i + U_i$. Sensitivity analyses used complete ascertainment, all-unknown-as-absent, and all-unknown-as-present definitions to bound the effect of the unresolved outcome status.

Two nested investigator-defined sensitivity endpoints were evaluated. The expanded CKM burden added diagnosed dyslipidemia (BPQ080) to the four primary disease domains. The expanded CKM–renal burden also included nephronlithiasis (KIQ026). These extensions were treated as study-specific burden definitions and were kept distinct from the 2026 AHA/ACC/ADA/ASN CKM staging algorithm.

An exact CKM condition count from 0 to 4 was calculated only when all four primary disease domains were observed. A survey-weighted quasi-Poisson companion model with a log link estimated the mean ratio in the exact CKM condition count per one-point increase in the reproductive-history index excluding gestational diabetes, adjusted for age, family income-to-poverty ratio (PIR), PHQ-9 score, race and Hispanic origin, education, and marital status.

Guideline-aligned CKM staging

Guideline-aligned CKM stages 0–4 were evaluated separately from the primary binary endpoint among non-pregnant women with positive fasting-subsample weights and complete observable staging inputs. Staging followed the 2026 AHA/ACC/ADA/ASN clinical practice guidelines to the extent permitted by public-use NHANES data (Ndumele *et al.*, 2026).

Stage 1 factors included excess or dysfunctional adiposity, defined by body mass index, waist circumference, and prediabetes. Asian-specific thresholds were used for the BMI and waist circumference. Stage 2 factors included hypertension, hypertriglyceridemia, metabolic syndrome, type 2 diabetes, and moderate-to-high-risk of CKD. Hypertension was defined as a systolic blood pressure of at least 130 mmHg, diastolic blood pressure of at least 80 mmHg, or antihypertensive treatment. Diabetes was defined as diagnosed diabetes, glucose-lowering medication use, HbA1c ≥6.5%, or fasting glucose ≥126 mg/dL. Prediabetes was defined as HbA1c 5.7%–6.4% or fasting glucose 100–125 mg/dL in the absence of diabetes.

Kidney function was characterized using the 2021 race-free CKD-EPI creatinine equation (Inker *et al.*, 2021) as follows:

$$G_i = 142\min\left(\frac{S_{\mathrm{cr},i}}{0.7},1\right)^{-0.241}\max\left(\frac{S_{\mathrm{cr},i}}{0.7},1\right)^{-1.200}(0.9938)^{a_i}(1.012) \quad (2)$$

where G_i denote eGFR and a_i age in years for woman i, and serum creatinine $S_{\mathrm{cr},i}$ is expressed in mg/dL and G_i in mL/min/1.73 m². Albuminuria categories were defined from urine albumin-creatinine ratio as A1 (<30 mg/g), A2 (30–<300 mg/g), and A3 (≥300 mg/g), and eGFR categories followed KDIGO G1–G5 thresholds. Moderate- to high-risk and very-high-risk CKD were mapped to the risk categories used by the 2026 CKM guideline.

For participants aged 30–44 years with complete supported inputs, the 10-year total CVD risk was estimated using the AHA PREVENT-CVD equation (Khan *et al.*, 2024). Inputs were restricted to the validated calculator ranges and were not imputed for the staging. Observable stage 3 was assigned for very-high-risk CKD or PREVENT-CVD risk ≥20%, consistent with the guideline's risk-equivalent criteria. Direct coronary artery calcium, ankle-brachial index, cardiac biomarkers, and echocardiographic measures were unavailable in the public-use data; therefore, the stage-3 estimate was explicitly interpreted as observable guideline-aligned stage 3, not as complete clinical ascertainment. Stage 4 requires clinical CVD. A survey-weighted proportional-odds companion model related the reproductive-history index excluding gestational diabetes to a higher CKM stage, adjusted for age, PIR, and PHQ-9 score.

Predictor specification and prevention of information leakage

Predictors were specified before model fitting and were restricted to variables that were not components of the primary or sensitivity CKM outcomes. The primary model included age, PIR, PHQ-9 score, number of pregnancies, a non-livebirth pregnancy proxy, race and Hispanic origin, educational attainment, marital status, ever-pregnant status, ever-livebirth status, any non-livebirth pregnancy proxy, gestational diabetes history, macrosomia history, high parity, and early or late age at first live birth.

The non-livebirth pregnancy variable was treated as a questionnaire-derived proxy rather than a miscarriage count. Questionnaire skip patterns were distinguished from true missingness: women who reported never having been pregnant received structural zeros for pregnancy-dependent variables, whereas unresolved responses remained missing. The current pregnancy was removed from the completed pregnancy count when ascertainable.

Hypertension, diabetes mellitus, dyslipidemia, CKD, clinical CVD, and nephrolithiasis were programmatically excluded from the predictor matrix and unsupervised phenotype analysis. This restriction prevented the outcome information from entering the predictor representation. The primary machine-learning specification used reproductive-history components rather than simultaneously entering a composite index and its constituent variables. A separate endpoint-comparison model used an adverse reproductive-history index excluding gestational diabetes so that diabetes did not appear in both the predictor and outcome domains.

PHQ-9 and missing predictor data

Each PHQ-9 item was considered valid only when coded 0–3. The PHQ-9 total score was calculated only for participants with all nine valid item responses; values used for refusal, uncertainty, or other non-responses were treated as missing.

Primary machine-learning preprocessing was estimated within each training fold and then applied to the corresponding validation fold. Continuous missing values were replaced by the survey-weighted median estimated from the training data. Operating under a missing-at-random (MAR) mechanism, this survey-weighted single-imputation approach was utilized to maintain complex design consistency while avoiding the prohibitive computational overhead and potential data leakage associated with nested multiple imputation (Hastie *et al.*, 2009; Little & Rubin, 2019). Categorical missing values were represented by an explicit missing level when observed during training, and validation categories absent from the training partition were mapped to the weighted modal training category. This procedure ensured that the imputation parameters, category definitions, and design matrices were learned without using validation outcomes or validation distributions.

Effective sample size and model-complexity assessment

Because unequal sampling weights reduce the information content relative to an equally weighted sample, Kish's effective sample size (n_{eff})was calculated as:

$$n_{\mathrm{eff}} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} \quad (3)$$

where n denotes the total number of individuals in the sample, and w_i represents the survey design weight assigned to participant i. The same expression was applied to the event weights to obtain an effective event count.

Model complexity was additionally examined with *pmsampsize* across candidate parameter counts and plausible C-statistics. This was used as a model-complexity diagnostic, not as a retrospective power calculation, and not as evidence that the fixed NHANES sample satisfied an a priori development sample-size requirement. The diagnostic was reported only as contextual information on model complexity and was interpreted according to the prediction-model sample-size criteria that emphasize shrinkage, optimism, and precision rather than a fixed events-per-variable rule (Riley *et al.*, 2019). It did not determine cohort inclusion or substitute for the internal validation.

The low number of primary CKM events motivated the use of constrained learner library, regulation penalization, limited nonlinear degrees of freedom, and predefined subgroup-support thresholds. Deep neural networks, Synthetic Minority Over-sampling Technique (SMOTE), random undersampling, and outcome-dependent class weighting were excluded from the primary probability models because they would add complexity or alter the probability target.

Survey-weighted empirical risk

Let n denote the total sample size, with $i = 1, \ldots, \mathrm{n}$ indexing the individual observations, w_i denoting the final sampling weight, X_i the predictor vector, $Y_i \in \{0,1\}$ the CKM endpoint, and $p_f(X_i)$ the probability produced by learner f. For learners that accepted case weights, normalized NHANES sampling weights entered the empirical loss so that fitting reflected the weighted target-population distribution. The weighted empirical risk is:

$$\hat{R}_w(f) = \frac{\sum_{i=1}^{n} w_i \, \ell\left(Y_i, p_f(X_i)\right)}{\sum_{i=1}^{n} w_i} \tag{4}$$

where ℓ denotes Bernoulli log loss or the corresponding learner loss. Within model fitting, weights were divided by their training-fold mean to preserve all relative weights while improving numerical conditioning. Design-based inference was obtained from the survey design and replicate-weight procedures described below; weight normalization did not alter relative sampling information.

Learner library and hyperparameter tuning

Hyperparameter optimization across all applicable learners relied on an exhaustive grid search strategy within the inner cross-validation folds**.** Specifically, a deterministic grid search evaluated all combinations of the mixing parameter and regularization paths for the elastic net, as well as the predefined maximum depth, learning rate, and child weight configurations for XGBoost and CatBoost. Survey-weighted spline logistic regression and the Explainable Boosting Machine utilized a predefined, fixed specification to serve as structurally constrained benchmarks, bypassing the need for a stochastic or heuristic search. Five learners were compared, as explained in the subsequent paragraphs.

Survey-weighted spline logistic regression analysis. The reference regression model used *svyglm* with a quasi-binomial family. Age, PIR, and PHQ-9 were modeled with natural cubic splines with three degrees of freedom; reproductive and categorical predictors were entered as predefined main effects. This model provides a transparent design-based benchmark.

Elastic-net logistic regression. The weighted elastic net was implemented using *glmnet*. The mixing parameter α was evaluated at 0, 0.25, 0.50, 0.75, and 1.00. A common 60-value λ path was log-spaced from 0.5 to 10^{-4}. Hyperparameters minimized the survey-weighted validation log loss in the grouped inner cross-validation. The final fit used the selected (α, λ) pair and normalized the training weights.

XGBoost. Weighted gradient boosting used a predefined grid of maximum tree depth of 2 or 3, learning rate of 0.02 or 0.05, and minimum child weight of 5 or 10. The subsample and column-subsample fractions were 0.85; L_2 regularization was 1.0 and L_1 regularization 0. The hyperparameters were selected using a weighted inner-fold log loss. The final fold-specific model used 700 boosting rounds.

CatBoost. CatBoost was called from R through *reticulate* and fitted with the observation weights. The candidate depth was 3 or 4 and the learning rate was 0.02 or 0.05, with L_2 leaf regularization 5. Inner tuning used 450 iterations and the fold-specific final fit used 700 iterations.

Explainable Boosting Machine. The EBM was fitted through InterpretML with observation weights. The primary EBM used main effects only (*interactions = 0*), 64 bins, a learning rate of 0.02, a minimum leaf size of 8,8 outer bags, and a 0.15 internal validation fraction. This predefined specification limited the model complexity while allowing nonlinear univariable effects.

Repeated survey-design–aware cross-validation

Internal validation used 20 repeats of five-fold cross-validation. Fold assignment was operated on PSU-within-stratum units so that all observations from the same PSU and stratum were assigned to the same validation fold. This preserves the cluster grouping and avoids dependence leakage between the training and validation samples. Candidate fold allocations were rejected if any validation fold lacked events or non-events. The same grouping principle was used in the 3-fold inner tuning loop. This approach was motivated by complex-survey cross-validation methods that account for stratification, clustering, and unequal selection probabilities (Wieczorek *et al.*, 2022).

For repeat r and outer fold k, all preprocessing, imputation, factor handling, spline construction, hyperparameter selection, and model fitting were restricted to the training partitions. Therefore, each participant therefore received exactly one out-of-fold prediction per learner in each repeat. A learner entered the model-comparison analysis only when complete out-of-fold prediction coverage was obtained for every participant across all the 20 repeats.

This resampling scheme estimates repeated out-of-sample performance within the observed survey while preserving the PSU grouping. This was not labeled as an external or temporal validation.

Cross-fitted probability recalibration

Raw out-of-fold probabilities were recalibrated separately within each model and outer-CV repeat using a second five-fold PSU-grouped cross-fitting layer. If $\hat{p}_i^{(0)}$ is a raw out-of-fold probability, the logistic recalibration model is:

$$\text{logit}\{\Pr(Y_i = 1)\} = \alpha + \beta \,\text{logit}\left\{\hat{p}_i^{(0)}\right\} \tag{5}$$

where α denotes the recalibration intercept, β is the calibration slope, Y_i represents outcome, and $\hat{p}_i^{(0)}$ is raw out-of-fold probability. Calibration-in-the-large was estimated separately by fitting an intercept-only recalibration with $\text{logit}\{\hat{p}_i^{(0)}\}$ included as an offset, thereby fixing its coefficient at 1. The mapping applied to an observation was always estimated using data that excluded that observation. Survey weights were used for the recalibration fit.

Logistic recalibration was preferred to isotonic regression because the event counts within the calibration training partitions were small. The primary model was selected by the mean survey-weighted calibrated log loss across the 20 outer-CV repeats, with the Brier score used as a secondary ordering criterion. The ROC AUC was not used as the primary selection criterion.

Predictive performance

The performance was evaluated at three levels: probability accuracy, discrimination, and calibration. All primary performance measures used survey weights corresponding to the target population.

The survey-weighted log loss, denoted $\mathcal{L}_w$, is expressed as follows:

$$\mathcal{L}_w = -\frac{\sum_{i=1}^{n} w_i \,[Y_i \log(\hat{p}_i) + (1 - Y_i)\log(1 - \hat{p}_i)]}{\sum_{i=1}^{n} w_i} \tag{6}$$

where n is the number of observations in the analytic sample, indexed by $i = 1,2, \ldots, n$; w_i is the final MEC examination sampling weight for observation i; $Y_i \in \{0,1\}$ is the observed CKM multimorbidity indicator, equal to 1 when the endpoint is present and 0 otherwise; $\hat{p}_i \in (0,1)$ is the out-of-fold predicted probability that $Y_i = 1$; and log denotes the natural logarithm. Normalization by $\sum_{i=1}^{n} w_i$ makes $\mathcal{L}_w$ a weighted mean that estimates the average log loss in the target population.

The survey-weighted Brier score, denoted B_w, is calculated as follows:

$$B_w = \frac{\sum_{i=1}^{n} w_i \,(Y_i - \hat{p}_i)^2}{\sum_{i=1}^{n} w_i} \tag{7}$$

where n, w_i, Y_i, and $\hat{p}_i$ are defined as in Eq. (6), and $(Y_i - \hat{p}_i)^2$ is the squared difference between the observed outcome and its predicted probability. $\mathcal{B}_w$ estimates the population mean squared prediction error and ranges from 0 (perfect prediction) to 1. Both are proper scoring rules; and lower values indicate better probability predictions.

Discrimination was summarized by survey-weighted ROC AUC and survey-weighted average precision, used as the empirical precision-recall AUC (PR AUC) estimator. For event set $\mathcal{E}$ and nonevent set $\mathcal{N}$, weighted AUC, denoted A_w, was computed as the weighted Mann–Whitney functional:

$$A_w = \frac{\sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{N}} w_i\, w_j \left[I(\hat{p}_i > \hat{p}_j) + \frac{1}{2} I(\hat{p}_i = \hat{p}_j)\right]}{\left(\sum_{i\in\mathcal{E}} w_i\right)\left(\sum_{j\in\mathcal{N}} w_j\right)} \tag{8}$$

where $\mathcal{E} = \{i: Y_i = 1\}$ is the set of observations with the CKM endpoint (events) and $\mathcal{N} = \{j: Y_j = 0\}$ is the set without it (nonevents); i and j index events and nonevents, respectively; w_i, w_j and $\hat{p}_i$, $\hat{p}_j$ are the corresponding sampling weights and predicted probabilities defined in Eq. (6); and $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 when its argument is true and 0 otherwise, so that tied predictions contribute to one-half. Each event–nonevent pair is weighted by $w_i w_j$, the estimated number of such pairs representing the target population. $\mathcal{A}_w$ estimates the probability that a randomly selected woman with CKM multimorbidity in the target population has a higher predicted probability than a randomly selected woman without it; a value of 0.5 indicates no discrimination and 1 indicates perfect discrimination.

Calibration was assessed using calibration-in-the-large, calibration slope, integrated calibration index (ICI), median absolute calibration error (E50), and 90th-percentile absolute calibration error (E90). The expected calibration error was retained as a supplementary bin-based summary. Design-based calibration intercept and slope were estimated with *svyglm*. A weighted penalized smooth calibration curve based on the logit of predicted probability was used to obtain ICI, E50, and E90. The distinction between discrimination and calibration followed established prediction-model evaluation guidance (Steyerberg *et al.*, 2010; Van Calster *et al.*, 2016; Van Calster *et al.*, 2019).

Repeat-specific performance was the primary internal-validation result. For each learner, the mean, standard deviation, and 2.5th and 97.5th percentiles of the 20 repeat-specific estimates were reported as measures of algorithmic variability; these repeat quantiles were not interpreted as sampling confidence intervals.

For secondary presentation and survey-inference calculations, participant-level repeat-averaged out-of-fold predictions (consensus predictions) were obtained by averaging only those predictions that were out of fold in their respective repeats. No apparent full-sample prediction was used to estimate the primary model performance.

Design-based uncertainty

Sampling uncertainty and algorithmic instability have been reported separately. For calibrated consensus predictions, log-loss and Brier-score contributions were treated as survey variables, and their means and standard errors were estimated using Taylor linearization under the NHANES design.

ROC AUC and PR AUC confidence intervals used a 500-replicate survey-bootstrap design created from the primary survey design, with the rank metrics recomputed within each replicate. Confidence limits used the replicate-design degrees of freedom. This approach avoided ordinary independent-sample DeLong intervals for data generated using a stratified, clustered probability sample.

Exploratory decision-curve analysis

Decision-curve analysis was specified in the final analysis plan as exploratory. At threshold t, survey-weighted net benefit $N_w(t)$ was:

$$N_w(t) = \frac{\sum_i w_i\ I(\hat{p}_i \geq t) Y_i}{\sum_i w_i} - \frac{\sum_i w_i\ I(\hat{p}_i \geq t)(1 - Y_i)}{\sum_i w_i} \frac{t}{1-t} \quad (9)$$

where $t \in (0,1)$ is the threshold probability at which a woman is classified as high risk; w_i, Y_i, and $\hat{p}_i$ are the sampling weight, observed CKM multimorbidity indicator, and predicted probability for observation i, as defined for Eq. (6); and $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 when its argument is true and 0 otherwise, so that $\mathbb{I}(\hat{p}_i \geq t) = 1$ identifies observations classified as high risk.

The first term is the weighted proportion of the target population classified as true positives, and the second is the weighted proportion classified as false positives, multiplied by the threshold odds $t/(1-t)$, which expresses the harm of a false positive relative to the benefit of a true positive implied by choosing threshold t. The reference strategies were:

$$N_w^{\text{all}}(t) = \bar{Y}_w - (1 - \bar{Y}_w)\frac{t}{1-t}, \qquad N_w^{\text{none}}(t) = 0,$$

$$\bar{Y}_w = \frac{\sum_i w_i\, Y_i}{\sum_i w_i}$$

where $N_w^{\text{all}}(t)$ and $N_w^{\text{none}}(t)$ are the net benefits of classifying all and no women as high risk, respectively, and $\bar{Y}_w$ is the survey-weighted outcome prevalence. The net benefit was evaluated for thresholds from 0.01 to 0.20 in increments of 0.005. Treat-all and treat-none strategies were used as references. Because the analysis was cross-sectional and lacked external or temporal validation, net benefit was interpreted as a threshold-dependent classification summary rather than evidence of clinical effectiveness or deployment readiness (Van Calster *et al.*, 2018).

Subgroup performance

Prespecified subgroup analyses considered age group, race and Hispanic origin, and PIR category. A subgroup estimate was reported only when the subgroup contained at least 50 observations, at least eight events, and at least 20 non-events. Kish effective sample size was also reported. These support thresholds were used to prevent unstable or uninterpretable subgroup performance estimates from being presented as precise evidence.

Model attribution

XGBoost was refitted on the complete analytic sample only after all internal-validation estimates were finalized. Exact TreeSHAP contributions were obtained with *predcontrib = TRUE* and aggregated using NHANES sampling weights. For feature j, the global importance summary was the survey-weighted mean absolute SHAP contribution. The full-data refit was used solely for model attribution and was never used to replace out-of-fold performance estimates. TreeSHAP values were interpreted as contributions to the fitted prediction function, rather than causal effects or biological mechanisms (Lundberg *et al.*, 2020).

Predictor-only secondary phenotyping

Phenotype analysis was secondary and descriptive. It excluded all components of the primary and expanded CKM outcomes before dimension reduction. Reproductive-history, socioeconomic status, age, and PHQ-9 variables were imputed using survey-weighted medians for this unsupervised analysis.

Let z_i be the standardized predictor vector for participant i. The weighted covariance matrix used for the principal component analysis was:

$$\Sigma_w = \frac{\sum_{i=1}^n w_i\, z_i z_i^{\top}}{\sum_{i=1}^n w_i} \quad (10)$$

where d is the number of predictors, n is the number of participants, w_i is the sampling weight for participant i, and $\mathbf{z}_i^{\top}$ denotes the transpose of $\mathbf{z}_i$. Because each predictor was standardized to weighted mean 0 and weighted standard deviation 1, $\boldsymbol{\Sigma}_w$ is the $d \times d$ survey-weighted correlation matrix of the predictors. Principal-component scores were obtained from the eigendecomposition of $\boldsymbol{\Sigma}_w$:

$$\boldsymbol{\Sigma}_w \mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0, \qquad \mathbf{u}_i = \mathbf{V}_q^{\top} \mathbf{z}_i$$

where λ_k and $\mathbf{v}_k$ are the kth eigenvalue and unit-length eigenvector, $\mathbf{V}_q = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$ is the $d \times q$ matrix of the first q eigenvectors, respectively; and $\mathbf{u}_i \in \mathbb{R}^q$ is the vector of retained principal-component scores for participant i. The number of retained components was the smallest number explaining at least 80% of the weighted variance, subject to a minimum of two and a maximum of eight components:

$$q = \min\{8,\ \max(2,\ q_{0.80})\},$$

$$q_{0.80} = \min\left\{m : \frac{\sum_{k=1}^m \lambda_k}{\sum_{k=1}^d \lambda_k} \geq 0.80\right\}$$

Weighted k-means then minimized:

$$W_K = \sum_{i=1}^n w_i \left\| u_i - \mu_{c(i)} \right\|^2 \quad (11)$$

over cluster assignments and centroids, where K is the number of clusters, $c(i) \in \{1, \ldots, K\}$ is the cluster to which participant i is assigned, $\boldsymbol{\mu}_{c(i)}$ is the centroid of that cluster, and $\|\cdot\|$ denotes the Euclidean norm. At the optimum, each centroid is the survey-weighted mean of the scores in its cluster:

$$\boldsymbol{\mu}_k = \frac{\sum_{i:c(i)=k} w_i\, \mathbf{u}_i}{\sum_{i:c(i)=k} w_i}, \qquad k = 1, \ldots, K$$

Candidate values of K from 2 to 6 were evaluated. The selected solution maximized the weighted average silhouette width among partitions in which every cluster had a Kish effective size of at least 30, where the effective size of cluster k is:

$$n_{\text{eff},k} = \frac{\left(\sum_{i:c(i)=k} w_i\right)^2}{\sum_{i:c(i)=k} w_i^2}$$

ties favored smaller K. CKM outcome prevalence was calcu-

lated only after cluster assignment and had no role in phenotype construction.

Survey-plasmode simulation benchmark

A survey-plasmode simulation was used to evaluate the operating characteristics of alternative validation strategies under a known finite-population truth. Empirical covariate vectors were resampled from the NHANES analytic data with a probability proportional to MEC sampling weights, preserving realistic predictor distributions and correlations. Each finite population contained 40,000 individuals arranged in 20 strata, four PSUs per stratum, and 500 individuals per PSU.

Outcomes were generated using a nonlinear logistic model with a PSU-level random effects. Let A_i, H_i, R_i, and Q_i denote standardized age, log-transformed PHQ-9 score, PIR, and pregnancy count, respectively. The data-generating model was as follows:

$$\begin{aligned}\operatorname{logit}(p_i^*) = a + 0.55A_i + 0.28A_i^2 + 0.30H_i - 0.22R_i + 0.22Q_i \\ + 0.35G_i + 0.28L_i + 0.25A_iH_i + b_{c(i)} \qquad (12)\end{aligned}$$

where G_i denotes gestational-diabetes history, L_i is any nonlivebirth pregnancy proxy, and $b_{c(i)} \sim N\left(0, \sigma_b^2\right)$ is the PSU random effect. The intercept a was numerically calibrated to target population prevalence of 2.0%, 6.6%, or 10.0%. The cluster dependence was varied with $\sigma_b = 0$ or 0.35.

Informative unequal-probability sampling was induced through the following:

$$s_i = \exp(\gamma A_i - 0.35\gamma R_i + 0.20\gamma H_i) \qquad (13)$$

where A_i, R_i, and H_i denote the age, PIR, and PHQ-9 terms for individual i as defined in Eq. (12), and $\gamma \geq 0$ governs the strength of informative selection and hence the variability of the sampling weights ($\gamma = 0$ would yield equal selection probabilities). Older women, women with lower PIR, and women with higher PHQ-9 scores therefore received higher selection scores; because these covariates also entered the outcome model in Eq. (12), selection was informative with respect to the outcome.

The within-PSU inclusion probabilities and simulation weights were as follows:

$$\pi_i = \min\{0.95, \max(0.002, c_{hc}s_i)\}, \qquad w_i^{\text{sim}} = \frac{1}{0.5\,\pi_i} \qquad (14)$$

where π_i is the conditional probability that individual i is sampled given that its PSU is selected, c_{hc} is the scaling constant for PSU c in stratum h, and w_i^{sim} is the simulation design weight, equal to the inverse of the unconditional two-stage inclusion probability $0.5\,\pi_i$. The truncation bounds of 0.002 and 0.95 exclude zero and near-certain inclusion and restrict w_i^{sim} to between approximately 2.1 and 1,000. We set $\gamma = 0.20$ for low and $\gamma = 0.70$ for high weight variability. Within each stratum, two of four PSUs were selected without replacement; the factor 0.5 in Eq. (14) is therefore the PSU inclusion probability.

To obtain an expected sample size of approximately 40 individuals per selected PSU after truncation, the scaling constant c_{hc} in Eq. (14) was determined separately for each PSU as the root of the estimating equation:

$$g_{hc}(c) = \sum_{i \in \mathcal{U}_{hc}} \min\{0.95,\ \max(0.002,\ c\,s_i)\} = 40 \qquad (14a)$$

where $\mathcal{U}_{hc}$ denotes the set of 500 individuals in PSU c of stratum h, $\kappa \geq 0$ is a candidate value of the scaling constant, and $g_{hc}(\kappa)$ is the resulting expected number of sampled individuals in that PSU. The function $g_{hc}(\kappa)$ is continuous and nondecreasing in κ, with boundary values:

$$g_{hc}(0) = 500 \times 0.002 = 1,$$

$$\lim_{c \to \infty} g_{hc}(c) = 500 \times 0.95 = 475$$

and is strictly increasing whenever at least one unit lies strictly between the truncation bounds. Because the target value of 40 lies within this range, Eq. (14a) admits a unique solution, $\kappa = c_{hc}$, which was obtained numerically using [Brent's method, as implemented in the R function uniroot, with a convergence tolerance of 10^{-8}]. As s_i is a function of covariates alone, c_{hc} depends on γ and on the covariate composition of each PSU, but is invariant to the target outcome prevalence.

The intercept a in Eq. (12) was calibrated by the same root-finding procedure to satisfy:

$$\frac{1}{N}\sum_{i=1}^{N} p_i^*(a) = \rho, \qquad N = 40{,}000,$$

$$\rho \in \{0.020,\ 0.066,\ 0.100\} \qquad (12a)$$

where N denotes the finite-population size, ρ the target outcome prevalence, and $p_i^*(a)$ the outcome probability of individual i from Eq. (12), written as a function of a with all other coefficients held fixed

The scenario grid crossed three outcome prevalence, two levels of weight variability, and two levels of cluster dependence, producing 12 scenarios. Each scenario included 500 Monte Carlo replicates, for 6,000 completed survey samples.

Four validation strategies were compared with the same fixed elastic-net learner ($\alpha = 0.5, \lambda = 0.01$) to isolate the effect of the validation method: (1) naive outcome-stratified random cross-validation without weights; (2) weighted random cross-validation; (3) weighted PSU-grouped survey-design–aware cross-validation; and (4) weighted PSU-grouped cross-validation followed by cross-fitted logistic recalibration. The log loss, Brier score, ROC AUC, and PR AUC were evaluated against the performance of the corresponding fitted model in the complete finite population.

For each performance metric $M \in$ {log loss, Brier score, ROC AUC, PR AUC} and Monte Carlo replicate $r = 1, \ldots, R$, with $R = 500$ replicates per scenario, the replicate-specific estimation error was $d_r = M_r - M_r^*$, where M_r is the estimate of M produced by a given validation strategy in replicate r, and M_r^* is the corresponding finite-population truth, that is, the performance of the model fitted in replicate r evaluated over all N individuals of the finite population. Monte Carlo bias B_M was:

$$B_M = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{M}_r - M_r^*\right) \tag{15}$$

Positive values of B_M indicate that a validation strategy overestimates M on average, and negative values indicate that it underestimates M. The root mean squared error R_M was:

$$R_M = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{M}_r - M_r^*\right)^2\right]^{1/2} \tag{16}$$

which combines systematic and random estimation errors, because $R_M^2 = B_M^2 + R^{-1}\sum_{r=1}^{R}(d_r - B_M)^2$. For replicate r, let $[L_r, U_r]$ denote the nominal 95% confidence interval for M produced by a given validation strategy. Empirical 95% interval coverage C_M was:

$$C_M = \frac{1}{R}\sum_{r=1}^{R} I\left(L_r \le M_r^* \le U_r\right) \tag{17}$$

where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 when the interval contains the finite-population truth and 0 otherwise. C_M is thus the proportion of replicates whose interval covers M_r^*, and a valid interval procedure yields C_M close to 0.95. With $R = 500$, the Monte Carlo standard error of a coverage estimate near 0.95 is approximately $\sqrt{0.95 \times 0.05/500} \approx 0.010$, so values between approximately 0.93 and 0.97 are consistent with the nominal coverage.

Monte Carlo standard errors were calculated from the between-replicate variation of the corresponding simulation estimators. Proper-score confidence intervals used ordinary sampling formulas only for the naive validation strategy; weighted strategies used a simulated complex-survey design. Simulation failures and non-finite summaries were logged, and the final run required 500 unique finite replicates per scenario.

Sensitivity analyses

Sensitivity analyses were used to address four sources of analytical uncertainty. *First*, endpoint definition was examined using the primary four-domain CKM definition, expanded CKM burden, expanded CKM–renal burden, a strict disease phenotype containing diabetes, CKD, and clinical CVD, complete ascertainment, and two extreme outcome-missingness bounds. *Second*, the exact 0–4 primary CKM condition count was modeled separately from the binary threshold. The companion count model assumes a log-linear mean structure with variance proportional to the mean:

$$\log\mu_i = \mathbf{x}_i^{\top}\boldsymbol{\beta}, \qquad \mathrm{Var}(Y_i \mid \mathbf{x}_i) = \phi\,\mu_i, \qquad \mathrm{MR} = \exp(\beta_{\mathrm{RH}})$$

where Y_i denotes the exact number of CKM conditions (range, 0–4), $\mathbf{x}_i$ the covariate vector, μ_i the conditional mean, and MR the adjusted mean ratio associated with a one-point increase in the reproductive-history index. The dispersion parameter ϕ was estimated as the weighted mean of the squared Pearson residuals:

$$\hat{\phi} = \frac{\sum_i w_i\, r_i^2}{\sum_i w_i}, \qquad r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

where w_i denotes the survey weight. Because the standard errors were derived from the design-based sandwich (Taylor-linearization) variance estimator, which remains consistent under the misspecification of the variance function, inference on the mean ratio does not depend on the value of ϕ. The estimate $\hat{\phi}$ is reported to characterize the degree of departure from the Poisson assumption of equidispersion ($\phi = 1$). *Third*, the guideline-aligned CKM stage was analyzed independently in the nonpregnant fasting subsample with the correct fasting weight and an explicit audit of unavailable stage-3 measurements. *Fourth*, a separate multiple-imputation module was retained as an optional missing-predictor sensitivity for transparent reproductive index regression. It was not used to tune, select, or evaluate primary machine-learning models.

Data analysis and reproducibility

Analyses were conducted in R 4.5.2 with Python 3.12.13 called through *reticulate* (random seed, 20260903). Five learners were evaluated using 20 repeats of 5-fold whole-PSU cross-validation. The uncertainty in ROC AUC and PR AUC used 500 survey-bootstrap replicates. The survey-plasmode benchmark comprised 12 scenarios with 500 Monte Carlo replicates each (6,000 total). The software versions, computational environments, input-file hashes, cross-validation checkpoints, simulation checkpoints, and SHA-256 output hashes were retained to support reproducibility.

The complete R/Python analysis code, environment specifications, and reproducibility materials are archived in Figshare (Adetunji, 2026, DOI: 10.6084/m9.figshare.33437179). Prediction-model reporting and risk-of-bias considerations were informed by TRIPOD+AI and PROBAST+AI (Collins *et al.*, 2024; Moons *et al.*, 2025), and the calibration assessment followed the established prediction-model guidance (Steyerberg *et al.*, 2010; Van Calster *et al.*, 2016; Van Calster *et al.*, 2019).

Results

Study population and CKM burden

The primary analytic cohort comprised 1,743 women aged 20–44 years, of whom 44 met the four-component CKM multimorbidity definition. The survey-weighted prevalence was 1.97% (95% CI, 1.30%–2.65%). **Figure 1A** shows the analytic flow, and Figure 1B shows the ascertainment of the four constituent disease domains and the partially identified multimorbidity endpoint. All 18 required NHANES files were loaded and hashed, and complete out-of-fold predictions were obtained for all 1,743 participants in each learner and cross-validation repeat (**Figure 1**).

Women with CKM multimorbidity were older than those below the threshold (weighted mean age 39.1 vs. 31.7 years; $P < 0.001$), had higher PHQ-9 scores (5.7 vs. 3.6; $P = 0.005$), and reported more pregnancies (2.7 vs. 1.9; $P = 0.002$). A history of gestational and high parity were also more common in the CKM group, although several reproductive history contrasts had limited precision because the primary endpoint contained only 44 events. The full design-weighted characteristics are presented in **Table 1**.

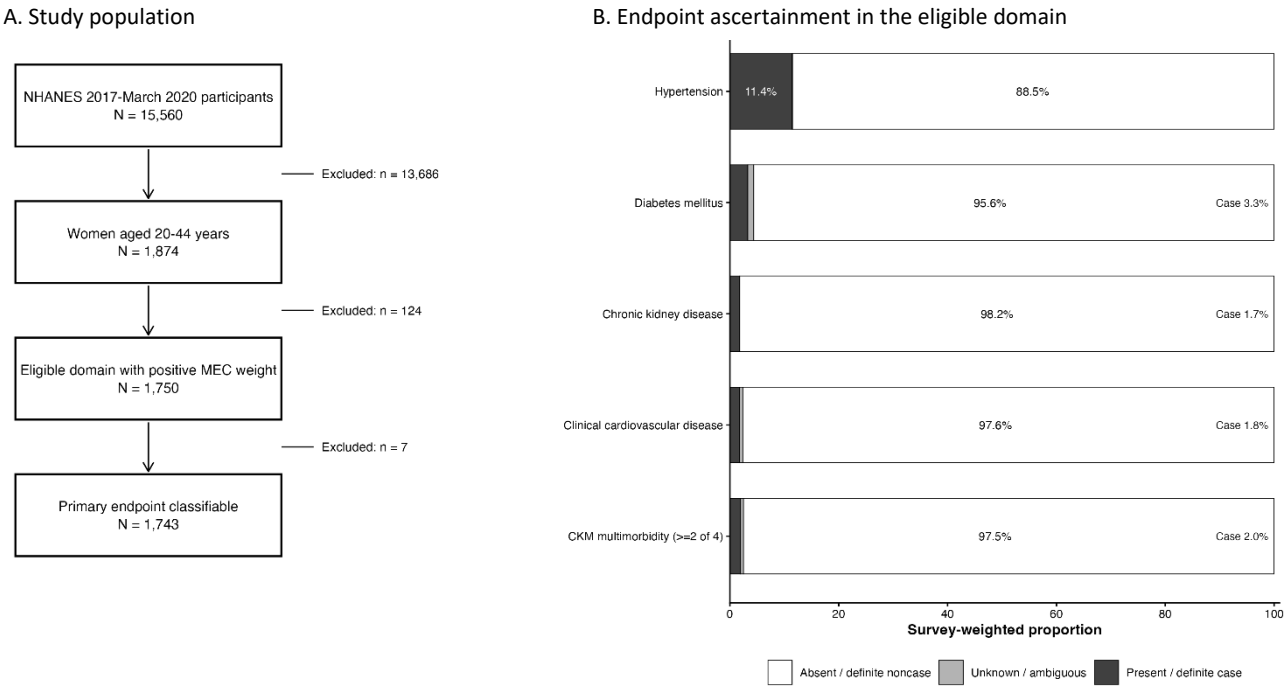


Figure 1 Study population and CKM endpoint ascertainment. (**A**) Flow from the NHANES 2017–March 2020 source population to women aged 20–44 years and the primary endpoint-classifiable cohort. (**B**) Survey-weighted ascertainment of hypertension, diabetes mellitus, CKD, clinical CVD, and the composite CKM multimorbidity endpoints. Unknown responses were preserved rather than automatically coded as disease absent.

Table 1 Characteristics of the analytic study population

Characteristic	Overall	No CKM multimorbidity	CKM multimorbidity	*P* value
Unweighted sample, n	1,743	1,699	44	-
Age, years, weighted mean (SE)	31.8 (0.2)	31.7 (0.2)	39.1 (0.4)	<0.001
Family income-to-poverty ratio, weighted mean (SE)	2.77 (0.10)	2.78 (0.10)	2.14 (0.42)	0.138
PHQ-9 score, weighted mean (SE)	3.7 (0.1)	3.6 (0.1)	5.7 (0.8)	0.005
Number of pregnancies, weighted mean (SE)	1.9 (0.1)	1.9 (0.1)	2.7 (0.3)	0.002
Non-livebirth pregnancy proxy, weighted mean (SE)	0.5 (0.0)	0.5 (0.0)	0.7 (0.2)	0.262
Gestational diabetes history, yes, weighted %	6.6	6.0	33.6	Descriptive
High parity, yes, weighted %	18.8	18.5	36.3	Descriptive
Any non-livebirth pregnancy proxy, yes, weighted %	26.9	26.8	33.8	Descriptive

Values are survey-weighted means (SE) or percentages unless otherwise stated. The *P* values were design-based and descriptive. CKM, cardiovascular-kidney-metabolic; PHQ-9, Patient Health Questionnaire-9; SE, standard error.

Model discrimination, probability accuracy, and calibration

The elastic net was selected as the primary learner because it minimized the mean calibrated survey-weighted log loss across the 20 repeated design-aware cross-validation runs. The mean log loss was 0.089 (SD 0.002), Brier score was 0.019, survey-bootstrap ROC AUC was 0.797 (95% CI, 0.765–0.830), PR AUC was 0.062 (0.023–0.100), mean calibration slope was 0.90 (SD 0.06), and mean ICI was 0.007. Given the weighted event prevalence of approximately 0.020, the PR AUC was substantially above the prevalence-defined no-skill reference while remaining appropriately modest for rare outcomes (**Table 2**).

CatBoost and XGBoost had similar Brier scores and slightly lower ROC-AUCs, whereas EBM had lower discrimination. The survey spline logistic benchmark had a consensus

ROC AUC of 0.574 after calibrated repeated cross-validation and showed an unstable repeat-specific calibration slope, illustrating the limits of an unregularized, parameter-rich benchmark with only 44 events. The model comparisons are presented in **Table 2**. The weighted ROC and precision–recall curves are shown in **Figure 2A** and **2B**, respectively.

The calibration curves demonstrated close agreement between the predicted and observed probabilities across the low absolute-risk range for the leading models as shown in **Figure 3A**. For the elastic-net consensus prediction, calibration-in-the-large was approximately zero, the calibration slope was 1.03, and while the repeatedCV mean slope was 0.90. The exploratory survey-weighted decision curves are shown in **Figure 3B**. These curves describe threshold-dependent classification trade-offs and were not used to infer the clinical effectiveness.

Endpoint hierarchy and guideline-aligned staging

The nested endpoint analysis showed the expected increase in burden as additional risk domains were included. Expanded CKM burden, which included dyslipidemia, was observed in 102 patients and had a weighted prevalence of 5.15% (95% CI, 3.8%–6.5%). Expanded CKM–renal burden, which additionally incorporated nephrolithiasis, included 138 cases and had aprevalence of 7.79% (6.2%–9.4%). In contrast, the strict disease phenotype defined by at least two of diabetes, CKD, and clinical CVD had 11 events and a prevalence of 0.45% (**Table 3**). Each one-point increase in the reproduc-

Table 2 Repeated design-aware cross-validated model performance

Model	Log loss, mean (SD)	Brier score, mean (SD)	ROC AUC (95% CI)	PR AUC (95% CI)	Calibration slope, mean (SD)	ICI, mean (SD)
Elastic net	0.089 (0.002)	0.019 (0.000)	0.797 (0.765–0.830)	0.062 (0.023–0.100)	0.90 (0.06)	0.007 (0.001)
CatBoost	0.090 (0.002)	0.019 (0.000)	0.760 (0.676–0.844)	0.067 (0.006–0.129)	0.87 (0.05)	0.005 (0.002)
XGBoost	0.091 (0.002)	0.019 (0.000)	0.755 (0.693–0.818)	0.067 (0.023–0.112)	0.86 (0.07)	0.005 (0.002)
EBM	0.093 (0.002)	0.019 (0.000)	0.723 (0.674–0.772)	0.048 (0.021–0.075)	0.81 (0.10)	0.006 (0.002)
Survey spline logistic	0.100 (0.003)	0.019 (0.000)	0.574 (0.474–0.674)	0.031 (0.013–0.049)	−0.37 (0.85)	0.006 (0.003)

The models were ordered by the calibrated survey-weighted log loss. The ROC AUC and PR AUC intervals used survey-bootstrap replicate weights for consensus out-of-fold predictions; other summaries described repeat-to-repeat internal-validation variability. AUC, area under the curve; CI, confidence interval; EBM, Explainable Boosting Machine; ICI, integrated calibration index; PR, precision-recall; ROC, receiver operating characteristic; SD, standard deviation.

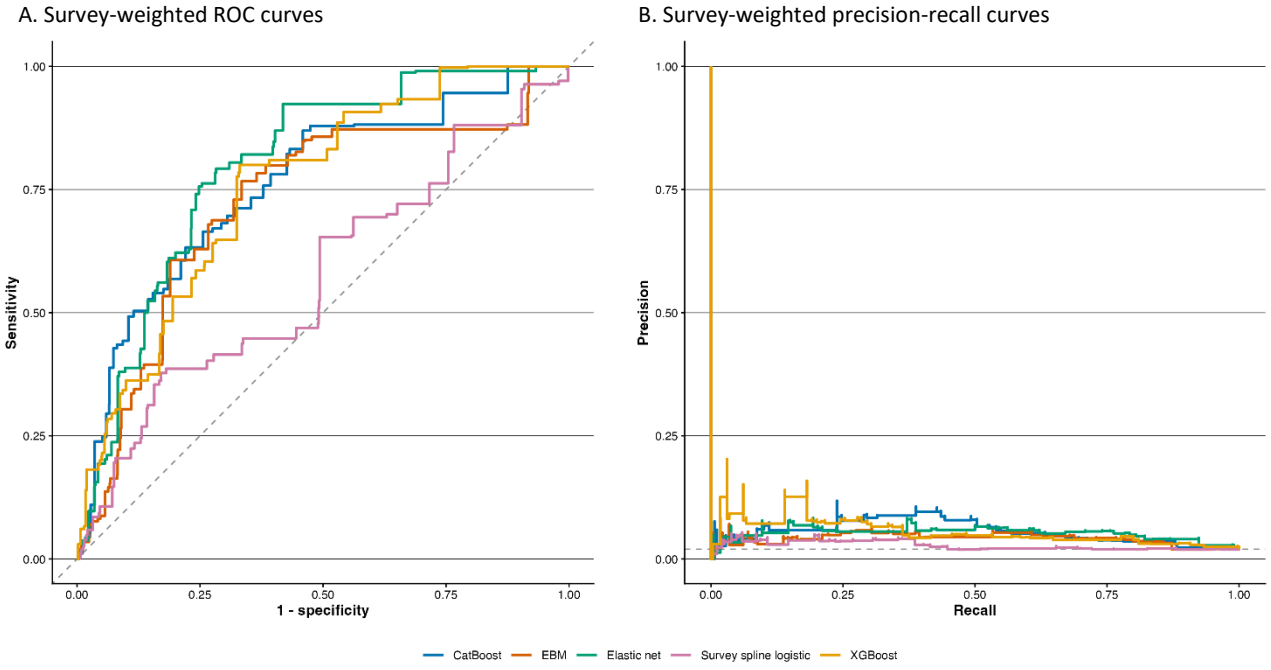


Figure 2 Survey-weighted discrimination of the CKM multimorbidity models. (**A**) ROC curves from consensus-calibrated out-of-fold predictions. (**B**) Precision–recall curves with weighted outcome prevalence shown as the no-skill reference. Model evaluation used survey weights and design-aware, repeated cross-validation.

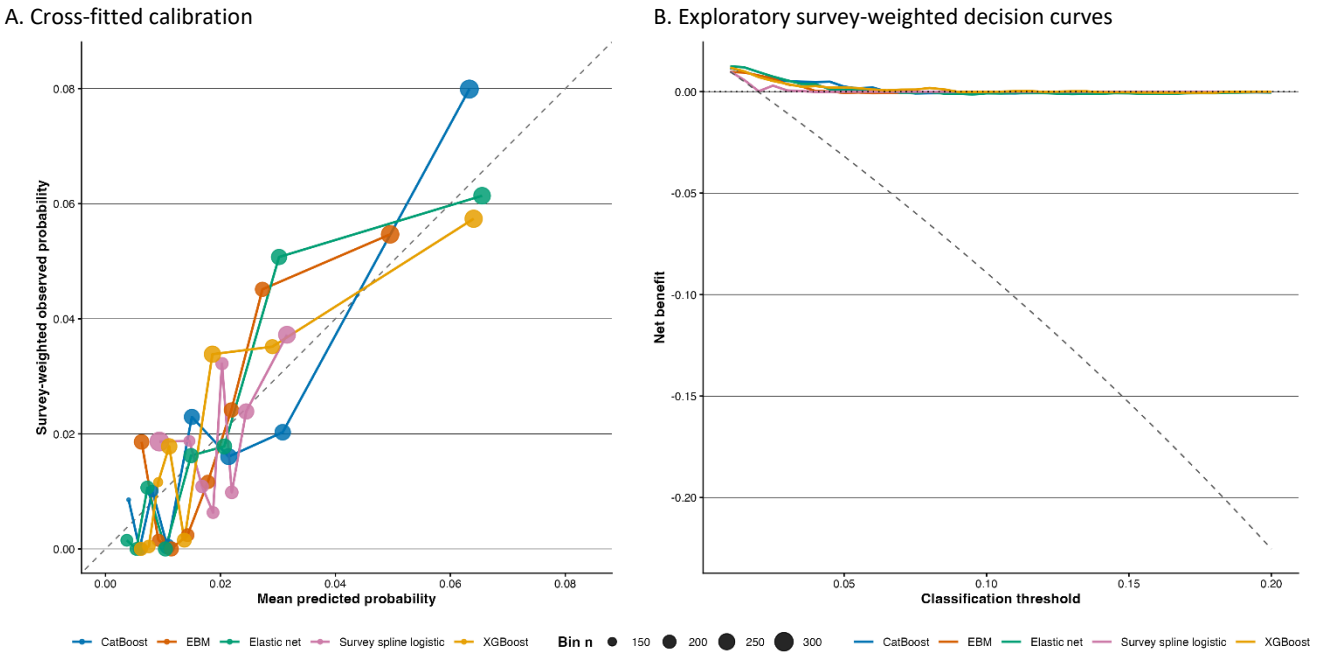


Figure 3 Probability calibration and exploratory decision curves. (**A**) Survey-weighted observed versus predicted CKM probabilities after design-aware cross-fitted calibration. (**B**) Exploratory survey-weighted decision curves across the prespecified classification thresholds. The analysis was concurrent and internally validated, the and net benefit was interpreted as a statistical threshold summary rather than prospective clinical utility.

tive-history index was associated with an adjusted mean ratio of 1.02 (95% CI, 0.82–1.26) for the number of CKM conditions. The estimated dispersion parameter was $\phi\hat{}$=0.98, indicating a minimal departure from equidispersion.

Complete ascertainment of all four primary disease components was available for 1,716 female participants. Among them, 84.5% had zero primary CKM conditions, 13.7% had one, 1.6% had two, and 0.2% had three or four primary CKM conditions. The weighted mean count was 0.18 cells/μL. Coding every ambiguous component as absent produced a prevalence of 2.0%, whereas coding every ambiguous component as present produced 2.5%, indicating limited movement across the extreme missingness bounds (**Table 3**).

Guideline-aligned CKM staging was available for 593 nonpregnant women in the fasting subsample. The weighted stage prevalence was 22.8% at stage 0, 33.1% at stage 1, 41.7% at stage 2, 0% at observable stage 3, and 2.4% at stage 4; therefore, stages 2–4 comprised 44.1% of the classifiable fasting subsample. The absence of observed stage 3 reflects both the young age range and the absence of direct subclinical-CVD measurements in the public-use NHANES. The primary CKM phenotype, expanded definitions, exact count, and staging results are reported in **Table 3** to clarify their distinct meanings.

Attribution, subgroup performance, and predictor-only phenotypes

In the full-data XGBoost explanation layer, age had the largest survey-weighted mean absolute TreeSHAP contribution, followed by the PHQ-9 score and income-to-poverty ratio. The number of pregnancies, history of gestational diabetes, and non-livebirth pregnancy proxy had smaller contributions. **Figure 4A** shown these model attributions. While these TreeSHAP values quantify the dependence of the fitted XGBoost function on measured predictors without identifying causal pathways, they crucially reflect mathematical feature importance within the model's specific architecture, representing statistical attribution rather than true biological or causal risk factors.

Prespecified subgroup performance was reported only when event-support thresholds were met (**Figure 4B**). For the elastic-net model, the ROC AUC was 0.805 among non-Hispanic White women and 0.733 among non-Hispanic Black women; estimates in other racial and ethnic groups were suppressed when event support was insufficient. Across age strata with adequate support, the ROC AUC was 0.729 at ages 35–39 and 0.544 at ages 40–44, illustrating greater uncertainty and weaker ranking within the oldest restricted age band (**Figure 4**).

The secondary predictor-only phenotype analysis yielded four descriptive profiles. The phenotype characterized by a universal history of gestational diabetes had the highest weighted CKM prevalence (10.0%), whereas the youngest low-reproductive-burden phenotype had a prevalence below 1%. Because outcome components were excluded before dimension reduction and clustering, these contrasts represent post-hoc outcome frequencies across independently constructed descriptive profiles. Detailed profiles are reported in the Supplementary Material **(Table S1** and **Figure S1)**.

Survey-plasmode benchmark

The 12-scenario, 6,000-replicate survey-plasmode experiment separated the influences of weighting, resampling structure, and calibration. Across scenarios, naive random cross-validation had a mean absolute log-loss bias of 0.0539 and a mean 95% log-loss interval coverage of 30.2%. Under

Table 3 CKM endpoint hierarchy, exact condition count, and guideline-aligned stage

Outcome	Analysis, n	Events/ category, n	Survey-weighted estimate (95% CI)	Adjusted reproductive-history index association	*p*-value
Primary CKM: ≥2 of hypertension, diabetes, CKD, CVD	1,743	44	2.0% (1.3–2.6)	OR 0.74 (0.42–1.29)	0.185
Expanded CKM: primary + dyslipidemia	1,738	102	5.1% (3.8–6.5)	OR 0.73 (0.48–1.12)	0.104
Expanded CKM–renal: expanded + kidney stones	1,735	138	7.8% (6.2–9.4)	OR 0.73 (0.49–1.09)	0.087
Strict disease phenotype: ≥2 of diabetes, CKD, CVD	1,746	11	0.4% (0.3–0.6)	OR 0.65 (0.18–2.35)	0.367
Primary, complete ascertainment	1,716	42	1.8% (1.4–2.3)	OR 0.80 (0.47–1.35)	0.266
Primary, missing coded absent	1,750	44	2.0% (1.3–2.6)	OR 0.75 (0.43–1.31)	0.197
Primary, missing coded present	1,750	51	2.5% (1.8–3.2)	OR 0.75 (0.39–1.46)	0.261
Exact count = 0	1,716	1,428	84.5% (82.0–86.9)	-	-
Exact count = 1	1,716	246	13.7% (11.3–16.0)	-	-
Exact count = 2	1,716	34	1.6% (1.1–2.1)	-	-
Exact count = 3–4	1,716	8	0.2%	-	-
CKM stage 0	593	-	22.8% (17.7–27.8)	-	-
CKM stage 1	593	-	33.1% (28.4–37.9)	-	-
CKM stage 2	593	-	41.7% (37.1–46.2)	-	-
CKM stage 3, observable criteria	593	0	0.0%	-	-
CKM stage 4	593	-	2.4% (1.6–3.3)	-	-

Expanded phenotypes are investigator-defined, nested sensitivity outcomes. Guideline stages were analyzed separately in the nonpregnant fasting subsample. Stage 3 used observable risk-equivalent criteria because direct subclinical CVD measurements were unavailable. Endpoint-comparison models used the reproductive-history index, excluding gestational diabetes, and were adjusted for age, income-to-poverty ratio, PHQ-9 score, race and Hispanic origin, education, and marital status. CI, confidence interval; CKD, chronic kidney disease; CKM, cardiovascular-kidney-metabolic; CVD, cardiovascular disease; OR, odds ratio; PHQ-9, Patient Health Questionnaire-9.

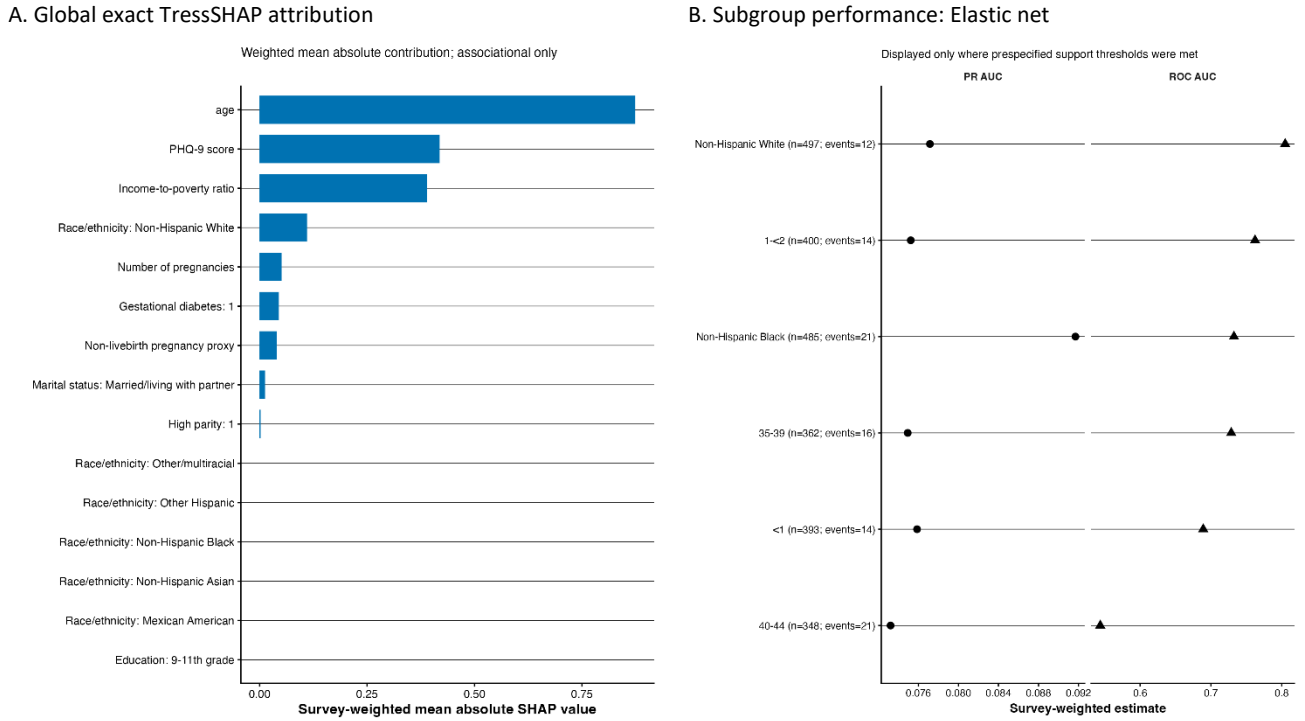


Figure 4 Model attribution and subgroup performance. (**A**) Survey-weighted exact TreeSHAP contributions from the final XGBoost explanation model; the attribution is associational. (**B**) Elastic-net subgroup performance displayed only where prespecified sample and event thresholds were satisfied.

high weight variability and 2% target prevalence, naive log-loss coverage was as low as 0.4%.

Weighting sharply reduced the performance estimation bias. Weighted whole-PSU design-aware cross-validation had a mean absolute log-loss bias of 0.0018 and a mean coverage of 90.4%. Adding cross-fitted calibration produced a mean log-loss coverage of 95.0% and a mean Brier-score coverage of 95.2%, with a mean absolute Brier bias of 0.0005. Therefore, the calibrated design-aware pipeline most closely recovered the nominal proper-score interval coverage over the full scenario grid, whereas weighted random and weighted design-aware cross-validation without calibration retained somewhat subnominal average coverage.

Figure 5A Shows log-loss bias relative to the finite-population truth, and **Figure 5B** shows the empirical 95% interval coverage. Scenario-specific bias, RMSE, Monte Carlo standard error, and coverage are provided in the Supplementary Material (**Table S2**).

Discussion

The principal contribution is a design-consistent, calibration-first statistical learning framework for the analysis of rare cardiovascular-kidney-metabolic multimorbidity using complex probability sampling. CKM multimorbidity, defined as the coexistence of at least two of hypertension, diabetes mellitus, chronic kidney disease, and clinical cardiovascular disease, had a survey-weighted prevalence of 1.97% among U.S. women aged 20–44 years. Although this baseline prevalence is lower than that in the broader U.S. adult population, where advanced CKM stages reach approximately 15% and climb to over 55% among adults aged 65 and older, these findings indicate that a substantial portion of reproductive-aged women are progressing along the early disease continuum, highlighting an important window for early clinical screening and preventive intervention before irreversible organ damage develops (Aggarwal *et al.*, 2024; Zhu *et al.*, 2024). In the fasting subsample, 44.1% of classifiable women were already in guideline-aligned CKM stages 2–4. These estimates distinguish investigator-defined multimorbidity phenotypes from guideline-aligned CKM staging along the disease continuum (Ndumele *et al.*, 2026). Stage describes the position on the disease continuum, whereas multimorbidity describes the accumulated disease burden. This distinction is clinically important because prevention begins before the establishment of multimorbidity. Moving from four major CKM disease domains to a definition that also included dyslipidemia more than doubled the estimated prevalence to 5.15%, and the further inclusion of nephrolithiasis increased it to 7.79%. This shows how the prevalence multimorbidity depends on the specific constituent conditions used to operationalize the definition; therefore, investigators should maintain distinct operational definitions to ensure cross-study reproducibility.

Predictive modeling within a complex probability sample is a finite-population estimation task rather than an algorithm selection exercise. Because the learner was held fixed across-validation schemes, the bias and undercoverage under naive cross-validation are attributable to the validation

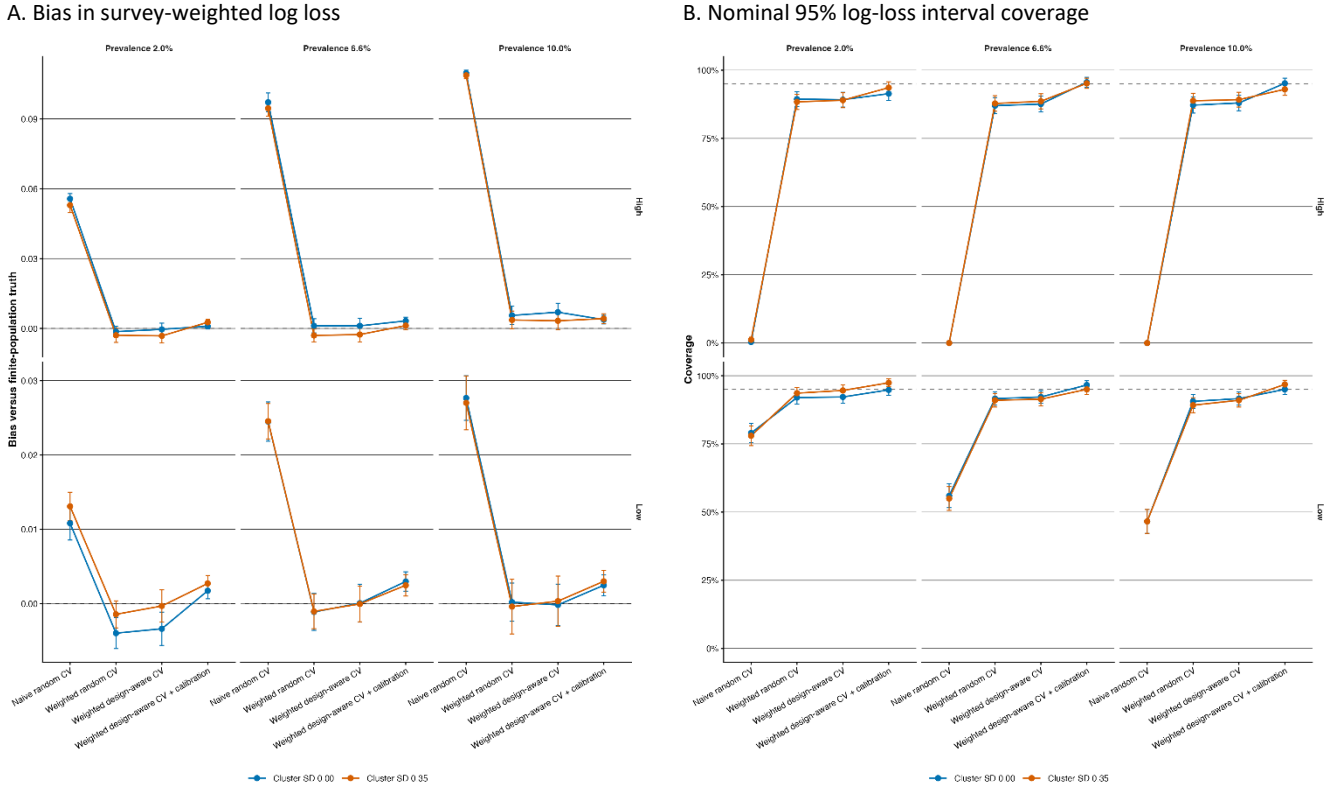


Figure 5 Survey-plasmode benchmark of the validation strategies. (**A**) Bias in survey-weighted log loss relative to the known finite-population truth across prevalence, weight-variability, and cluster-dependence scenarios. Error bars denote Monte Carlo uncertainty. (**B**) Empirical coverage of nominal 95% log-loss intervals; the horizontal reference denotes 95% coverage.

framework and not to model complexity. This supports Wieczorek *et al.* (2022), who argued that cross-validation for design-based inference must respect stratification, clustering, and unequal inclusion probabilities, and MacNell *et al.* (2023), who showed in NHANES that ignoring survey weights inflated the performance metrics of gradient-boosting models against population-appropriate values. Survey weights govern target population representation; primary sampling units determine the complex dependence structure; survey-weighted proper scoring rules quantify prediction error on the population scale; out-of-sample recalibration preserves population-level probability calibration; and design-based uncertainty procedures support valid variance estimation of performance metrics. Valid population-level inferences requires all five to refer to the same target population.

For rare outcomes, discrimination and calibration answer different questions. An ROC AUC of 0.797 indicates acceptable discrimination; however, the PR AUC of 0.062 must be interpreted relative to the survey-weighted baseline prevalence of approximately 0.020. Precision–recall analysis is informative under extreme class imbalance because ROC curves can appear overly optimistic, even when positive predictions exhibit low precision (Saito & Rehmsmeier, 2015). Calibration evaluates a distinct performance domain: whether the predicted probability corresponds to the observed finite-population event frequency. Valid clinical decision support requires agreement between the probability scale and empirical outcomes, rather than merely correct ranking (Van Calster *et al.*, 2016). The relative superiority of the elastic net model over the unpenalized spline-logistic benchmark is attributable to regularization rather than a difference in the functional form. The elastic net effectively regularizes the parameter space through simultaneous ridge and lasso penalties. This adaptive shrinkage mitigates the overfitting and structural instability that commonly affect unpenalized, parameter-rich models under severe event scarcity (Hastie *et al.*, 2009; Riley *et al.*, 2019).

Reproductive history provides additional context for this early disease interval. Women meeting the primary endpoint were older, had higher PHQ-9 scores, and reported more pregnancies; gestational diabetes and high parity were also more frequent. Longitudinal evidence links pregnancy history with later cardiovascular and renal health. Barrett *et al.* (2020), in a meta-analysis of more than 5.7 million women, reported an association between preeclampsia and gestational hypertension with chronic and end-stage kidney disease. Garovic *et al.* (2020) reported a higher risk of stroke, coronary artery disease, chronic kidney disease, and multimorbidity after hypertensive disorders of pregnancy. The present analysis adds to the existing literature by showing that a composite reproductive-history score did not exhibit a monotonic association with CKM risk. Individual reproductive events have heterogeneous predictive values because hypertensive disorders, gestational diabetes, parity, pregnancy loss, and fetal growth abnormalities reflect distinct pathways. Future studies should model these components separately rather than as an additive index and could use pregnancy history as an entry point for postpartum CKM screening (Parikh *et al.*, 2021; O'Kelly *et al.*, 2022).

Methodological strengths of this study include consistent alignment of the estimand, survey design features, resampling, calibration, and variance estimation; fold-contained pre-processing; and tri-state ascertainment that preserved indeterminate responses rather than recoding them as disease absence. The plasmode benchmark, based on a known finite-population truth, supports the relevance of the framework beyond CKM.

However, this study had several limitations. *First*, the primary endpoint had few events, limiting the stability of subgroup estimates. We addressed this with a parsimonious predictor set, penalization, restricted spline flexibility, repeated cross-validation, and minimum-support reporting rules. *Second*, NHANES is cross-sectional; therefore the model classifies prevalent diseases and cannot establish the temporal order of reproductive exposures and CKM onset. *Third*, several disease domains relied on self-reported clinician diagnoses; tri-state ascertainment and extreme missing-as-absent and missing-as-present bounds limited, but did not eliminate, misclassification. *Fourth*, staging was confined to non-pregnant fasting participants and lacked direct measures of subclinical cardiovascular disease, such as coronary artery calcium, ankle-brachial index, cardiac biomarkers, and echocardiography. *Fifth*, this analysis remains subject to unmeasured confounding from unobserved parameters, including genetic predisposition and unmeasured lifestyle factors; thus, the top predictors identified by the models represent prognostic risk markers rather than definitive causal pathways. External validation in geographically and temporally distinct cohorts is required before clinical decision-support.

Future studies utilizing this framework should follow a clear methodological roadmap focused on extending design-consistent validation to longitudinal datasets and integrating high-dimensional genomic features. Additionally, deploying open-source software libraries will help standardize calibration-based machine learning across complex surveys.

Conclusion

Among U.S. women aged 20–44 years, the survey-weighted prevalence of concurrent CKM multimorbidity was 1.97%, whereas 44.1% of classifiable women in the fasting subsample were in guideline-aligned CKM stages 2–4 years. A calibrated elastic-net model discriminated CKM multimorbidity with an ROC AUC of 0.797 (95% CI, 0.765–0.830). These estimates distinguish two related but distinct constructs: CKM multimorbidity reflects the coexistence of established cardiovascular, kidney, and metabolic diseases, whereas CKM staging classifies individuals according to their progression along the CKM continuum. The substantial prevalence of stages 2–4 suggests that metabolic risk factors and kidney disease are already common among reproductive-aged women before multimorbidity or clinical cardiovascular disease develops, supporting early risk identification and prevention.

For rare outcomes in complex probability samples, valid estimation of predictive performance requires design-consistent, calibration-first validation. In the survey-plasmode benchmark, nominal 95% log-loss intervals from naive random cross-validation achieved a mean coverage of only 30.2%, whereas survey-weighted, PSU-preserving validation with cross-fitted recalibration achieved 95.0% coverage for log loss and 95.2% for the Brier score. These results show that valid inferences require alignment among the target population, sampling design, validation scheme, calibration procedure, performance measures, and variance estimator. Sampling weights represent the target population; PSU-preserving resampling respects within-cluster dependence; proper scoring rules quantify probabilistic prediction errors; cross-fitted recalibration assesses and corrects calibration in held-out data; and design-based variance estimation quantifies uncertainty under the complex sampling design. Therefore, the predictive performance in complex surveys therefore depends on both the fitted model and the procedures used for validation and inference. This principle extends to any national probability sample when performance estimates are intended to be applied to a defined target population, and explicit reporting of these design elements should accompany such studies.

Acknowledgments The authors acknowledge the academic training environment in biostatistics and epidemiology provided through the the College of Health, Oregon State University, which contributed to the methodological foundation of this work. The authors also acknowledge the National Center for Health Statistics, Centers for Disease Control and Prevention, for making the National Health and Nutrition Examination Survey public-use data available for research.

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

Conflict of interest All authors declare no conflicts of interest related to this study.

Author contribution SAA: Conceptualization, data curation, formal analysis, investigation, methodology, software, supervision, validation, visualization, writing – original draft, and writing – review & editing. ROO: Investigation, clinical interpretation, writing – review and editing. TB: Investigation, clinical interpretation, and writing – review & editing. BA: Investigation, clinical interpretation, writing – review and editing. OAA: Project administration, study coordination, investigation, and writing – review & editing.

Availability of data and materials NHANES 2017–March 2020 pre-pandemic public-use data are available from the National Center for Health Statistics, Centers for Disease Control and Prevention. Analysis code, software environment specifications, and reproducibility materials are archived in Figshare, DOI: 10.6084/m9.figshare.33437179. The software record will be made publicly available upon the publication of the associated article.

Declaration of generative AI in writing process The authors independently architected and wrote all software code, statistical analyses, and manuscript text. OpenAI was consulted solely to troubleshoot isolated syntax and runtime errors during code debugging. All scholastic work was executed, verified, and validated by the authors, who take full responsibility for the final publication.

Received 4 September 2026 | **Revised** 7 November 2026 | **Accepted** 12 November 2026 | **Published online** 20 November 2026

Supplementary material

Supplementary data associated with this article (containing **Table S1**, **Table S2**, **Figure S1**, and **Figure S2**) are available online alongside the published version of this article.

References

Adetunji SA. 2026. svylearn: Design-consistent statistical learning, calibration, and inference for complex surveys. Figshare. Software. Version 1.0.0. DOI: 10.6084/m9.figshare.33437179.

Aggarwal R, Ostrominski JW, Vaduganathan M. 2024. Prevalence of cardiovascular-kidney-metabolic syndrome stages in US adults, 2011–2020. *JAMA*, 331(21): 1858–1860. DOI: 10.1001/jama.2024.6892.

Barrett PM, McCarthy FP, Kublickiene K, Cormican S, Judge C, Evans M, Kublickas M, Perry IJ, Stenvinkel P, Khashan AS. 2020. Adverse pregnancy outcomes and long-term maternal kidney disease: a systematic review and meta-analysis. *JAMA Network Open*, 3(2): e1920964. DOI: 10.1001/jamanetworkopen.2019.20964.

Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385: e078378. DOI: 10.1136/bmj-2023-078378.

Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. 2024. Developing clinical prediction models: a step-by-step guide. *BMJ*, 386: e078276. DOI: 10.1136/bmj-2023-078276.

Garovic VD, White WM, Vaughan L, Saiki M, Parashuram S, Garcia-Valencia O, Weissgerber TL, Milic N, Weaver A, Mielke MM. 2020. Incidence and long-term outcomes of hypertensive disorders of pregnancy. *Journal of the American College of Cardiology*, 75(18): 2323–2334. DOI: 10.1016/j.jacc.2020.03.028.

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York (NY): Springer Science & Business Media.

Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, Crews DC, Doria A, Estrella MM, Froissart M, Grams ME, Greene T, Grubb A, Gudnason V, Gutiérrez OM, Kalil R, Karger AB, Mauer M, Navis G, Nelson RG, Poggio ED, Rodby R, Rossing P, Rule AD, Selvin E, Seegmiller JC, Shlipak MG, Torres VE, Yang W, Ballew SH, Couture SJ, Powe NR, Levey AS; Chronic Kidney Disease Epidemiology Collaboration. 2021. New creatinine- and cystatin C-based equations to estimate GFR without race. *New England Journal of Medicine*, 385(19): 1737–1749. DOI: 10.1056/NEJMoa2102953.

Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, Blaha MJ, Carson AP, Chang AR, Ciemins E, Go AS, Gutierrez OM, Hwang SJ, Jassal SK, Kovesdy CP, Lloyd-Jones DM, Shlipak MG, Palaniappan LP, Sperling L, Virani SS, Tuttle K, Neeland IJ, Chow SL, Rangaswami J, Pencina MJ, Ndumele CE, Coresh J; Chronic Kidney Disease Prognosis Consortium and AHA CKM Science Advisory Group. 2024. Development and validation of the American Heart Association's PREVENT equations. *Circulation*, 149(6): 430–449. DOI: 10.1161/CIRCULATIONAHA.123.067626.

Kim H, Rogers P, Wang D. 2026. Machine learning models for chronic disease risks using complex survey data: the impact of sample weights. *Frontiers in Digital Health*, 8: 1834715. DOI: 10.3389/fdgth.2026.1834715.

Little RJ, Rubin DB. 2019. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken (NJ): John Wiley & Sons.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67. DOI: 10.1038/s42256-019-0138-9.

MacNell N, Feinstein L, Wilkerson J, Salo PM, Molsberry SA, Fessler MB, Thorne PS, Motsinger-Reif AA, Zeldin DC. 2023. Implementing machine learning methods with complex survey data: lessons learned on the impacts of accounting sampling weights in gradient boosting. *PLoS ONE*, 18(1): e0280387. DOI: 10.1371/journal.pone.0280387.

McNestry C, Killeen SL, Crowley RK, McAuliffe FM. 2023. Pregnancy complications and later life women's health. *Acta Obstetricia et Gynecologica Scandinavica*, 102(5): 523–531. DOI: 10.1111/aogs.14523.

NCHS [National Center for Health Statistics]. 2022. National Health and Nutrition Examination Survey, 2017–March 2020 pre-pandemic file: sample design, estimation, and analytic guidelines. Hyattsville (MD): National Center for Health Statistics. Vital and Health Statistics, Series 2, No. 190: 1–36. DOI: 10.15620/cdc:115434.

Ndumele CE, Neeland IJ, Tuttle KR, Chow SL, Mathew RO, Khan SS, Coresh J, Baker-Smith CM, Carnethon MR, Després JP, Ho JE, Joseph JJ, Kernan WN, Khera A, Kosiborod MN, Lekavich CL, Lewis EF, Lo KB, Ozkan B, Palaniappan LP, Patel SS, Pencina MJ, Powell-Wiley TM, Sperling LS, Virani SS, Wright JT, Singh RR, Elkind MSV, Rangaswami J; American Heart Association. 2023a. A synopsis of the evidence for the science and clinical management of cardiovascular-kidney-metabolic (CKM) syndrome: a scientific statement from the American Heart Association. *Circulation*, 148(20): 1636–1664. DOI: 10.1161/CIR.0000000000001186.

Ndumele CE, Rangaswami J, Chow SL, Neeland IJ, Tuttle KR, Khan SS, Coresh J, Mathew RO, Baker-Smith CM, Carnethon MR, Després JP, Ho JE, Joseph JJ, Kernan WN, Khera A, Kosiborod MN, Lekavich CL, Lewis EF, Lo KB, Ozkan B, Palaniappan LP, Patel SS, Pencina MJ, Powell-Wiley TM, Sperling LS, Virani SS, Wright JT, Singh RR, Elkind MSV; American Heart Association. 2023b. Cardiovascular-kidney-metabolic health: a presidential advisory from the American Heart Association. *Circulation*, 148(20): 1606–1635. DOI: 10.1161/CIR.0000000000001184.

Ndumele CE, Rodriguez F, Dixon DL, Khan SS, Mukherjee D, Bajaj M, Bangalore S, Bozkurt B, Breathett K, Clarke SL, de Boer IH, Ellison DH, Evangelista LS, Heffron SP, Kazi DS, Kulshreshtha A, Lingvay I, Low Wang CC, Mercado CA, Morton JM, Neeland IJ, Pagidipati N, Powell-Wiley TM, Rangaswami J, Rao G, Reza N, Saeed A, St Peter W, Starks JB, Sterling M, Talbot AW, Tran AH, Tuttle KR, VanWagner LB, Vest AR, Virani SS. 2026. 2026 AHA/ACC/ADA/ASN guideline for the prevention, detection, evaluation, and management of cardiovascular-kidney-metabolic syndrome: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, 154(4): e50–e158. DOI: 10.1161/CIR.0000000000001453.

O'Kelly AC, Michos ED, Shufelt CL, Vermunt JV, Minissian MB, Quesada O, Smith GN, Rich-Edwards JW, Garovic VD, El Khoudary SR, Honigberg MC. 2022. Pregnancy and reproductive risk factors for cardiovascular disease in women. *Circulation Research*, 130(4): 652–672. DOI: 10.1161/CIRCRESAHA.121.319895.

Okoth K, Chandan JS, Marshall T, Thangaratinam S, Thomas GN, Nirantharakumar K, Adderley NJ. 2020. Association between the reproductive health of young women and cardiovascular disease in later life: umbrella review. *BMJ*, 371: m3502. DOI: 10.1136/bmj.m3502.

Parikh NI, Gonzalez JM, Anderson CAM, Judd SE, Rexrode KM, Hlatky MA, Gunderson EP, Stuart JJ, Vaidya D. 2021. Adverse pregnancy outcomes and cardiovascular disease risk: unique opportunities for cardiovascular disease prevention in women: a scientific statement from the American Heart Association. *Circulation*, 143(18): e902–e916. DOI: 10.1161/CIR.0000000000000961.

Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, Collins GS. 2019. Minimum sample size for developing a multivariable prediction model: Part II—binary and time-to-event outcomes. *Statistics in Medicine*, 38(7): 1276–1296. DOI: 10.1002/sim.7992.

Riley RD, Collins GS, Kirton L, Snell KIE, Ensor J, Whittle R, Dhiman P, van Smeden M, Liu X, Alderman J, Nirantharakumar K, Manson-Whitton J, Westwood AJ, Cazier JB, Moons KGM, Martin GP, Sperrin M, Denniston AK, Harrell FE Jr, Archer L. 2025. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *BMJ*, 388: e080749. DOI: 10.1136/bmj-2024-080749.

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3): e0118432. DOI: 10.1371/journal.pone.0118432.

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1): 128–138. DOI: 10.1097/EDE.0b013e3181c30fb2.

Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. 2016. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74: 167–176. DOI: 10.1016/j.jclinepi.2015.12.005.

Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. 2018. Reporting and interpreting decision curve analysis: a guide for investigators. *European Urology*, 74(6): 796–804. DOI: 10.1016/j.eururo.2018.08.038.

Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; STRATOS Topic Group. 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17: 230. DOI: 10.1186/s12916-019-1466-7.

Wieczorek J, Guerin C, McMahon T. 2022. K-fold cross-validation for complex sample surveys. *Stat*, 11(1): e454. DOI: 10.1002/sta4.454.

Zhu R, Wang R, He J, Wang L, Chen H, Niu X, Sun Y, Guan Y, Gong Y, Zhang L, An P, Li K, Ren F, Xu W, Guo J. 2024. Prevalence of cardiovascular-kidney-metabolic syndrome stages by social determinants of health. *JAMA Network Open*, 7(11): e2445309. DOI: 10.1001/jamanetworkopen.2024.45309.