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

Supplementary material

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

Authors: Sunday A. Adetunji, Rhoda O. Oyewusi, Tomiisin Balogun, Bukola Adetunji, Oluwatimileyin A. Adetunji

Journal: Curr Biomed 5(1): 7–23, DOI: 10.29244/currbiomed.5.1.7

(**Table S1**, **Table S2**, **Figure S1**, and **Figure S2**)

Table S1 Predictor-only phenotype profiles

Phenotype	Unweighted (n)	Effective (n)	Age, weighted (mean)	PIR, weighted (mean)	PHQ-9, weighted (mean)	Pregnancies, weighted (mean)	GDM history, weighted (%)	High parity, weighted (%)	CKM prevalence, weighted (%)
Phenotype 1	689	370.3	33.9	2.60	3.1	2.2	0.0	24.6	1.3
Phenotype 2	452	238.0	27.0	3.07	3.8	0.0	0.0	0.0	0.9
Phenotype 3	127	62.6	35.2	2.53	4.5	3.5	100.0	43.8	10.0
Phenotype 4	475	263.0	34.1	2.65	3.7	3.6	0.0	27.8	2.2

Outcome components were excluded from the principal component analysis and clustering; CKM prevalence was calculated only after phenotype assignment. Effective n denotes the Kish effective sample size. CKM, cardiovascular-kidney-metabolic; GDM, gestational diabetes mellitus; PHQ-9, Patient Health Questionnaire-9; PIR, family income-poverty ratio.

Table S2 Survey-plasmode benchmark summary across scenarios

Validation strategy	Mean absolute log-loss bias	Mean 95% log-loss interval coverage	Mean absolute Brier-score bias	Mean 95% Brier-score interval coverage
Naive random CV	0.0539	30.2%	0.0189	28.7%
Survey-weighted random CV	0.0022	89.7%	0.0008	90.2%
Survey-weighted whole-PSU CV	0.0018	90.4%	0.0008	90.6%
Survey-weighted whole-PSU CV + cross-fitted recalibration	0.0026	95.0%	0.0005	95.2%

Values summarize 12 scenarios crossing target prevalences of 2.0%, 6.6%, and 10.0%, low versus high sampling-weight variability, and absent versus greater PSU-level dependence; each scenario included 500 Monte Carlo replicates (6,000 total). Coverage is the empirical coverage of nominal 95% intervals. CV, cross-validation; PSU, primary sampling units.

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

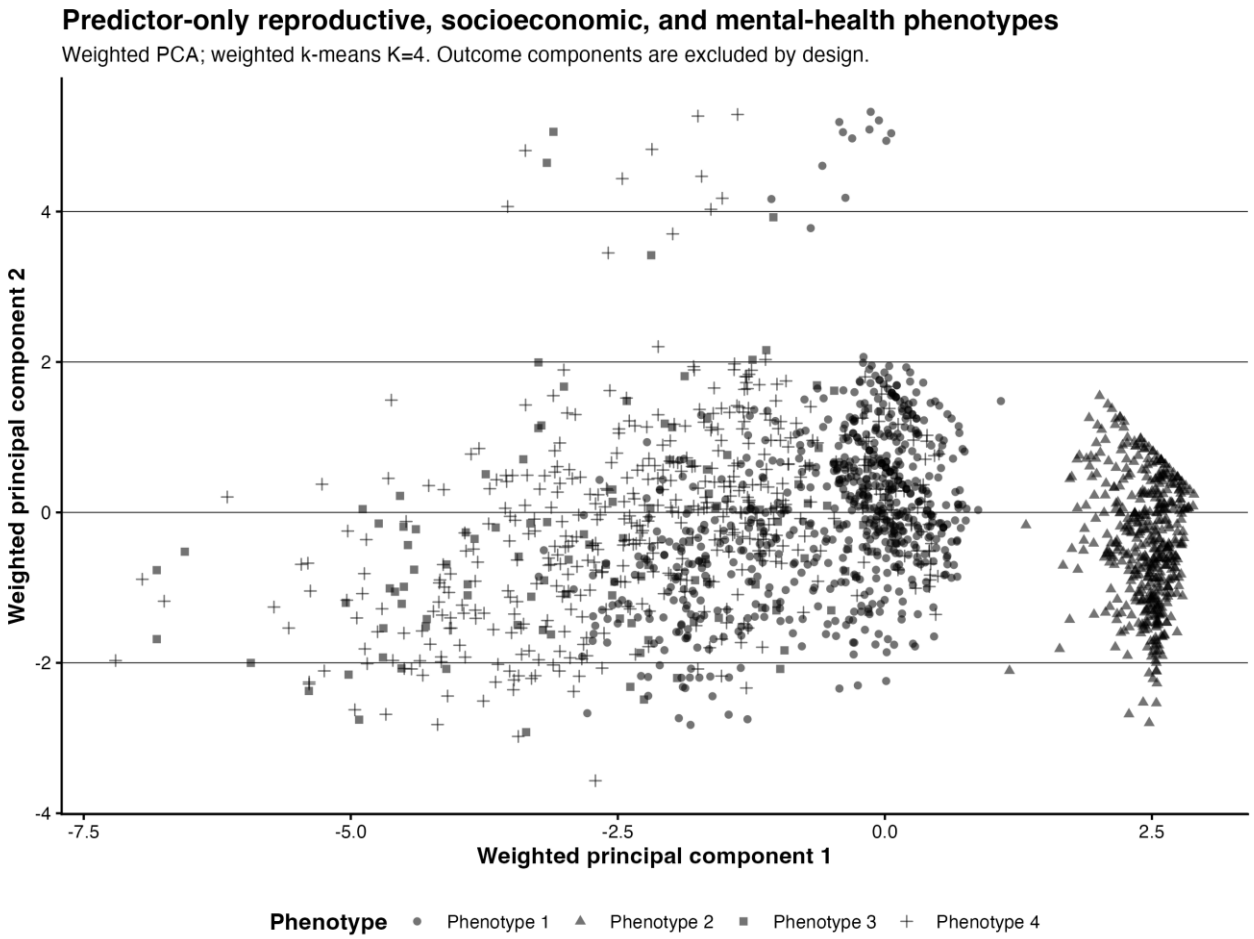


Figure S1 Predictor-only weighted phenotypes. Survey-weighted principal component analysis followed by weighted k-means clustering identified four descriptive phenotypes using reproductive-history, socioeconomic status, age, and PHQ-9 variables. All CKM outcome components were excluded before dimension reduction and clustering analysis. CKM prevalence was calculated only after phenotype assignment and did not contribute to cluster formation.

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

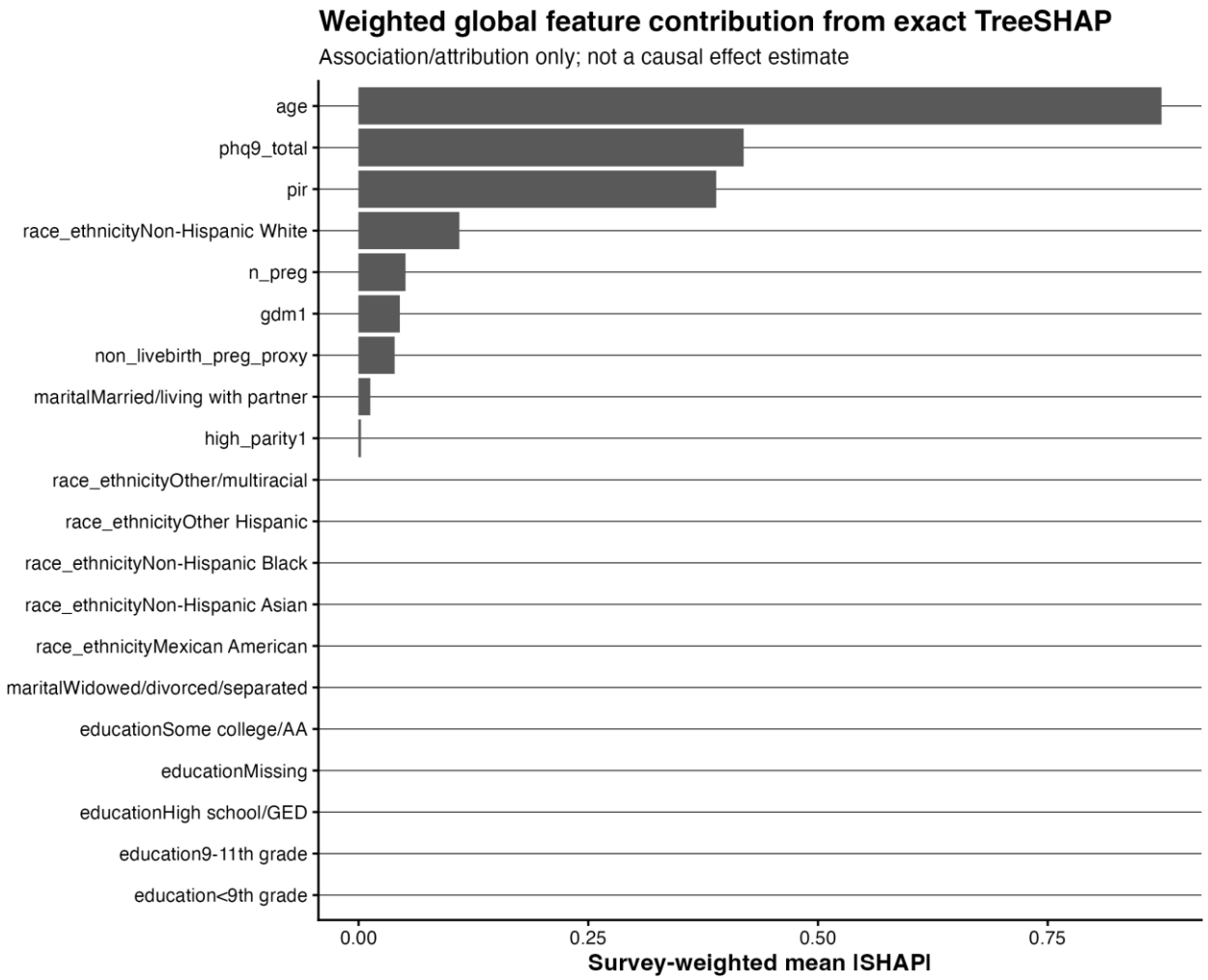


Figure S2 Exact TreeSHAP attribution for the XGBoost explanation model. The survey-weighted mean absolute TreeSHAP contributions summarize the relative contribution of each predictor to the fitted XGBoost prediction function. The XGBoost model was refitted to the complete analytic sample only after the internal-validation performance was finalized. TreeSHAP values represent model attribution and should not be interpreted as causal effects or biological mechanisms.