

Article

Not peer-reviewed version

Design-Aware Predictive and Causal Modeling of Cardiovascular Risk in Chronic Kidney Disease Using Penalized and Double Machine Learning Approaches

[Fernando Rojas](#)*, Axa Tapia, [Hilda Espinoza](#)

Posted Date: 9 April 2026

doi: 10.20944/preprints202604.0571.v1

Keywords: chronic kidney disease; cardiovascular risk; complex survey design; penalized regression; LASSO; double/debiased machine learning; two-stage residual inclusion; bootstrap inference; calibration; risk prediction



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Design-Aware Predictive and Causal Modeling of Cardiovascular Risk in Chronic Kidney Disease Using Penalized and Double Machine Learning Approaches

Fernando Rojas ^{1,2,*}, Axa Tapia ² and Hilda Espinoza ²

¹ Centro de Micro-Bio Innovación. Escuela de Nutrición y Dietética, Facultad de Farmacia, Universidad de Valparaíso, Chile
² Magíster en Análisis Clínico, Escuela de Química y Farmacia, Facultad de Farmacia, Universidad de Valparaíso, Chile
* Correspondence: fernando.rojas@uv.cl; Tel.: +56994389426

Abstract

We develop a design-aware mathematical framework that unifies penalized prediction and causal inference for finite populations observed through complex survey designs. The framework integrates survey-weighted pseudo-likelihoods, ℓ_1 -penalized estimation, Neyman-orthogonal moment functions, and a design-consistent bootstrap that resamples primary sampling units within strata. Methodologically, the contribution is an explicit pipeline that preserves design consistency while separating predictive associations from structural causal effects in high-dimensional, clustered data. We illustrate the framework using data from the Chilean National Health Survey (ENS) 2016–2017 to study the relationship between chronic kidney disease (CKD) and high cardiovascular (CV) risk. In the ENS adult population, the survey-weighted prevalence of CKD was 3.1% (95% CI: 2.4–3.8), and the prevalence of high CV risk was 23.9% (95% CI: 21.5–26.3). High CV risk was markedly more frequent among individuals with CKD than among those without CKD (90.9% versus 21.5%). Predictive and associational analyses combined survey-weighted penalized logistic regression (LASSO) with refitted unpenalized models. In conventional survey-weighted logistic regressions, CKD showed a strong associational relationship with high CV risk (odds ratio = 5.66; 95% CI: 2.71–11.82; $p < 0.001$), and effect sizes remained stable after LASSO-based variable selection. To assess causal relevance under confounding and potential endogeneity, we implemented two endogeneity-aware estimators: two-stage residual inclusion (2SRI) and Double/Debiased Machine Learning (DML). The DML estimator, defined as the primary causal estimand, reports the average treatment effect of CKD on the probability of high CV risk. After adjustment for age and major cardiometabolic comorbidities, the DML estimate was attenuated and statistically non-significant (average treatment effect = -0.094 ; 95% CI: $[-0.409, 0.220]$). The 2SRI approach yielded highly unstable odds-ratio estimates with wide confidence intervals, consistent with limited effective sample size and weak identification power. Simulation experiments under ENS-like complex sampling confirmed that naive predictive associations substantially overestimate the structural contribution of CKD under confounding, whereas orthogonalized estimators recover attenuated causal effects when identification holds. Overall, the results demonstrate a fundamental divergence between predictive relevance and causal importance in finite-population settings, underscoring the need for design-aware and endogeneity-robust methods in mathematical statistics and applied risk modeling.

Keywords: chronic kidney disease; cardiovascular risk; complex survey design; penalized regression; LASSO; double/debiased machine learning; two-stage residual inclusion; bootstrap inference; calibration; risk prediction

1. Introduction

Modern population analytics increasingly require methods that reconcile two distinct inferential goals—*prediction* and *causal estimation*—in data generated by complex sampling designs. In

multistage probability samples with stratification, clustering, and unequal inclusion probabilities, valid population-level statements hinge on design-consistent estimators, survey-aware performance metrics, and resampling schemes that mimic the sampling process [1–3]. At the same time, high-dimensional covariate sets and multicollinearity motivate penalized estimation and stability assessments grounded in pseudo-likelihood arguments and stochastic resampling [4,5]. When the scientific interest is causal—e.g., estimating a structural effect in the presence of confounding or endogeneity—orthogonalized estimators that leverage Neyman-orthogonal moments and cross-fitting provide a principled route to robust inference even when nuisance components are learned by flexible methods [6].

Recent advances have extended penalized methods to complex surveys, but most contributions address variable selection and penalty tuning without delivering an integrated pipeline that (i) evaluates prediction with survey-consistent discrimination and calibration, (ii) quantifies uncertainty via design-preserving bootstrap, and (iii) corrects endogeneity when estimating structural parameters. For example, Iparraguirre et al. [7] develop design-based cross-validation for LASSO in surveys, while McConville et al. [8] study LASSO and adaptive LASSO from a model-assisted, finite-population perspective; both strengthen the predictive regularization toolkit for complex designs, yet neither fully integrates endogeneity corrections or weighted predictive metrics. Likewise, Jasiak and Tuvaandorj [9] advance post-LASSO inference with weights mainly at the theoretical level, without embedding it into an end-to-end predictive-causal workflow.

Positioning relative to the existing literature.

The proposed framework extends several active lines of methodological research. Prior work on penalized estimation for complex surveys includes the design-based cross-validation and variable-selection developments of Iparraguirre et al. [7], the model-assisted LASSO formulations of McConville et al. [8], and the theoretical contributions of Jasiak and Tuvaandorj on penalized likelihood with survey weights [9]. These approaches primarily address prediction and regularization but do not incorporate formal treatment of endogeneity. Likewise, double/debiased machine learning (DML) [10] and recent advances in weighted orthogonal learners for causal identification [11] offer robust estimation tools, yet they do not provide an integrated pipeline that simultaneously addresses penalized survey-weighted prediction, design-consistent bootstrap inference, and endogeneity-aware causal estimation. To the best of our knowledge, no prior work combines these components into a unified framework supported by mathematical theory, simulation evidence, and application to real-world complex survey data.

Objective and contribution.

This paper proposes a unified, design-consistent framework that jointly addresses prediction and causality in complex survey data. The framework combines: (i) survey-weighted pseudo-likelihood estimation for finite-population targets, (ii) ℓ_1 -penalization (LASSO) to stabilize prediction under multicollinearity and limited effective sample size, (iii) design-consistent bootstrap (PSU-within-strata) for uncertainty and stability, and (iv) endogeneity-aware causal estimators—Two-Stage Residual Inclusion (2SRI) and Double/Debiased Machine Learning (DML)—implemented with survey weights and cross-fitting. Predictive performance is assessed using weighted AUC and Brier score, while structural effects are estimated through Neyman-orthogonal moments to mitigate regularization bias and confounding. Methodologically, the contribution is to deliver a coherent pipeline that makes explicit when predictive and causal conclusions diverge under complex sampling and how to report each with appropriate, design-aware uncertainty.

This paper makes four primary mathematical contributions within the study of penalized estimation and causal inference under complex survey designs:

1. Design-consistency of penalized survey-weighted estimators. We formalize a LASSO-penalized pseudo-likelihood estimator and establish its design-consistency under multistage sampling.

Theorem 1 characterizes convergence of the penalized estimator toward its finite-population target under appropriate regularity conditions.

2. Bootstrap consistency for design-based uncertainty quantification. We show that a stratified, PSU-level bootstrap consistently approximates the sampling distribution of the penalized estimator. This validates bootstrap inference for both predictive and causal parameters under complex sampling.
3. Operator-theoretic formulation of penalized estimation. We demonstrate that the survey-weighted LASSO estimator may be expressed as a stochastic proximal operator acting on the pseudo-score. This provides a unified operator-theoretic view linking optimization, survey sampling, and stochastic analysis.
4. Integration of orthogonal moments with survey weights. We extend double/debiased machine learning (DML) to complex surveys by incorporating sampling weights into nuisance estimation, cross-fitting at the PSU level, and orthogonal score construction. This ensures robustness of structural estimates to regularization bias and design-induced variability.

These components together form a coherent mathematical framework that connects penalized prediction, causal estimation, and design-based inference for finite populations.

Scope and illustration.

To illustrate the framework, we consider an application where predictive and causal goals are routinely conflated: the relationship between chronic kidney disease (CKD) and cardiovascular (CV) risk. CKD is epidemiologically linked to elevated CV risk and is often embedded in risk stratification, yet its *independent* structural contribution is sensitive to confounding and endogeneity; in such settings, naive predictive associations may not carry causal meaning [12,13]. We therefore use two complementary exercises. First, ENS-like simulations with known causal structure show how naive survey-weighted penalized models can overstate structural effects under confounding, whereas 2SRI/DML recover the target when identification holds. Second, we apply the pipeline to the Chilean National Health Survey (ENS) 2016–2017 to examine whether routinely collected biomarkers (hs-CRP, lipid fractions, hemoglobin, HbA1c, creatinine) improve prediction of high CV risk and how endogeneity-aware estimators alter the interpretation of the CKD effect in a finite population.

Organization.

Section 2 introduces the mathematical and statistical foundations (pseudo-likelihood under complex sampling, penalization, design-based bootstrap, and endogeneity corrections). Section 3 details the modeling pipeline—survey-weighted estimation, LASSO regularization, weighted performance metrics, and orthogonal/control-function estimators. Section 4 reports simulation and empirical results, focusing on the divergence between predictive and causal targets and on design-consistent uncertainty. Section 5 discusses implications, limitations, and avenues for future work. Finally, Section 6 summarizes the main insights and emphasizes the generalizability of the framework to other complex health surveys and application areas.

2. Motivation and Mathematical Background

Applied Motivation

The mathematical literature has recently emphasized the need for risk prediction models that account for non-independence among biomarkers, high-dimensionality, and survey sampling variability [14]. Penalized estimators such as Least Absolute Shrinkage and Selection Operator (LASSO) have demonstrated utility for feature selection under these constraints, particularly when effective sample size is reduced due to complex survey designs [15].

The physiological mechanisms underlying cardiovascular disease in CKD have been extensively described [16,17]. Inflammatory markers such as hs-CRP have shown predictive value for cardiovascular outcomes [18], although their interpretation in CKD remains challenging due to the presence of

uremic inflammation [19]. Oxidative stress is also implicated, with oxidized LDL linked to accelerated atherosclerosis [20]. Biomarkers such as cystatin C have proven especially informative, outperforming creatinine as predictors of cardiovascular and renal outcomes [21,22]. Beyond predictive performance, a central mathematical challenge is potential endogeneity between CKD status and cardiovascular risk—where CKD may influence cardiovascular outcomes, but cardiovascular events may also exacerbate CKD [23–25]. In econometric terms, endogeneity arises when regressors correlate with the error term, leading to biased estimates. To address this, we incorporate statistical techniques tailored for nonlinear survey-weighted models:

- **Control-function (Two-Stage Residual Inclusion, 2SRI):** For logistic regression with endogenous variables, we include residuals from a first-stage survey-weighted model as covariates in the second-stage outcome model. This method is consistent under nonlinearity and extends to complex designs [26,27].
- **Double/Debiased Machine Learning (DML):** We employ orthogonal estimators via Neyman-orthogonal moment conditions with sample splitting and penalized learners. This approach mitigates both regularization bias and endogeneity, and it can be adapted for survey designs to yield valid inference on the causal CKD parameter [10,28].

These methods preserve the weighted pseudo-likelihood framework and allow bootstrap-based uncertainty quantification, ensuring robust and interpretable inference on CKD effects under complex sampling.

Given the abundance of routine biomarkers collected in ENS 2016–2017 and the lack of CKD-specific predictive models in Chile, this study offers both mathematical and applied contributions.

A Stochastic Framework for Penalized Estimation under Complex Sampling

Let $\{(Y_i, X_i, w_i)\}_{i=1}^n$ denote observations drawn from a finite population $\mathcal{U} = \{1, \dots, N\}$ under a multistage, stratified sampling design. Each sampled unit carries a survey weight $w_i = 1/\pi_i$, where π_i is the inclusion probability. Only the sample is observed; the full population remains unobserved.

The survey-weighted pseudo-likelihood for parameter vector β is defined as

$$L(\beta) = \prod_{i=1}^n p_{\beta}(Y_i | X_i)^{w_i}, \quad (1)$$

where p_{β} denotes the logistic probability kernel. Taking logarithms yields the weighted pseudo-log-likelihood:

$$\ell(\beta) = \sum_{i=1}^n w_i \left[Y_i \log p_{\beta}(X_i) + (1 - Y_i) \log(1 - p_{\beta}(X_i)) \right]. \quad (2)$$

Under standard regularity conditions, expression (2) provides a design-consistent approximation to the finite-population log-likelihood.

To incorporate regularization and enable stochastic variable selection, we consider the LASSO-penalized objective

$$\mathcal{Q}_{\lambda}(\beta) = \ell(\beta) - \lambda \|\beta\|_1, \quad \lambda \geq 0, \quad (3)$$

whose maximizer satisfies

$$\hat{\beta}_{\lambda} = \arg \max_{\beta \in \mathbb{R}^p} \mathcal{Q}_{\lambda}(\beta). \quad (4)$$

This optimization defines a stochastic mapping

$$\hat{\beta}_{\lambda} = T_{\lambda}(\{(Y_i, X_i, w_i)\}_{i=1}^n), \quad (5)$$

whose randomness reflects both the sampling design and the underlying finite population.

To characterize the sampling distribution of $\hat{\beta}_\lambda$, we use a design-consistent bootstrap that resamples clusters and strata according to the ENS 2016–2017 design. This produces bootstrap replicates

$$\hat{\beta}_\lambda^{(b)}, \quad b = 1, \dots, B, \quad (6)$$

whose empirical distribution converges to that of $\hat{\beta}_\lambda$ under the complex sampling design.

This stochastic framework offers two key advantages:

1. It provides uncertainty quantification for penalized estimators under complex survey designs, where classical asymptotics may be unreliable.
2. It yields a finite-population interpretation of penalized pseudo-likelihood estimation, supporting inference on population-level cardiovascular risk mechanisms.

Thus, the proposed approach integrates survey theory, penalized estimation, and stochastic resampling into a unified mathematical framework.

Theorem: Design-Consistent Penalized Estimation in Finite Populations

Theorem 1. Let $\hat{\beta}_\lambda$ denote the LASSO-penalized pseudo-likelihood estimator obtained under a stratified multistage sampling design with inclusion probabilities $\{\pi_i\}$. Suppose:

- (A1) The sampling design satisfies design consistency for the pseudo-likelihood score.
- (A2) The logistic model is correctly specified for the finite population.
- (A3) The penalty parameter satisfies $\lambda_N \rightarrow 0$ and $\sqrt{n}\lambda_N \rightarrow \infty$.
- (A4) The number of predictors grows slowly with sample size: $p = o(n)$.

Then,

$$\hat{\beta}_\lambda \xrightarrow{P} \beta^*, \quad (7)$$

where β^* is the logistic minimizer at the population level.

Furthermore, the bootstrap estimator satisfies:

$$\sup_{t \in \mathbb{R}^p} \left| \Pr^* (\sqrt{n}(\hat{\beta}_\lambda^{(b)} - \hat{\beta}_\lambda) \leq t) - \Pr (\sqrt{n}(\hat{\beta}_\lambda - \beta^*) \leq t) \right| \rightarrow 0. \quad (8)$$

Sketch of proof. Assumption (A1) ensures convergence of the weighted pseudo-score to its finite-population counterpart. Assumption (A3) balances sparsity and asymptotic bias. Convexity of (3) yields consistency via M-estimation theory. Bootstrap consistency follows from resampling primary sampling units within strata, reproducing the sampling distribution of (4). $\square$

Stochastic Stability Paths

To assess robustness of penalized estimators under finite-population and design-induced randomness, consider a grid of tuning parameters $\Lambda = \{\lambda_1, \dots, \lambda_K\}$. For each λ_k and bootstrap replicate b , compute

$$\hat{\beta}^{(b)}(\lambda_k), \quad (9)$$

and define the selection probability for predictor j as

$$\hat{\pi}_j(\lambda_k) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(\hat{\beta}_j^{(b)}(\lambda_k) \neq 0). \quad (10)$$

A predictor is considered *stochastically stable* if

$$\hat{\pi}_j(\lambda_k) \approx 1 \quad \text{across a wide range of } \lambda_k. \quad (11)$$

These stability paths generalize classical LASSO solution paths by incorporating sampling variability and design-based uncertainty. Figure 1 summarizes the workflow from finite population to risk prediction, highlighting how bootstrap replicates feed into stability path estimation.

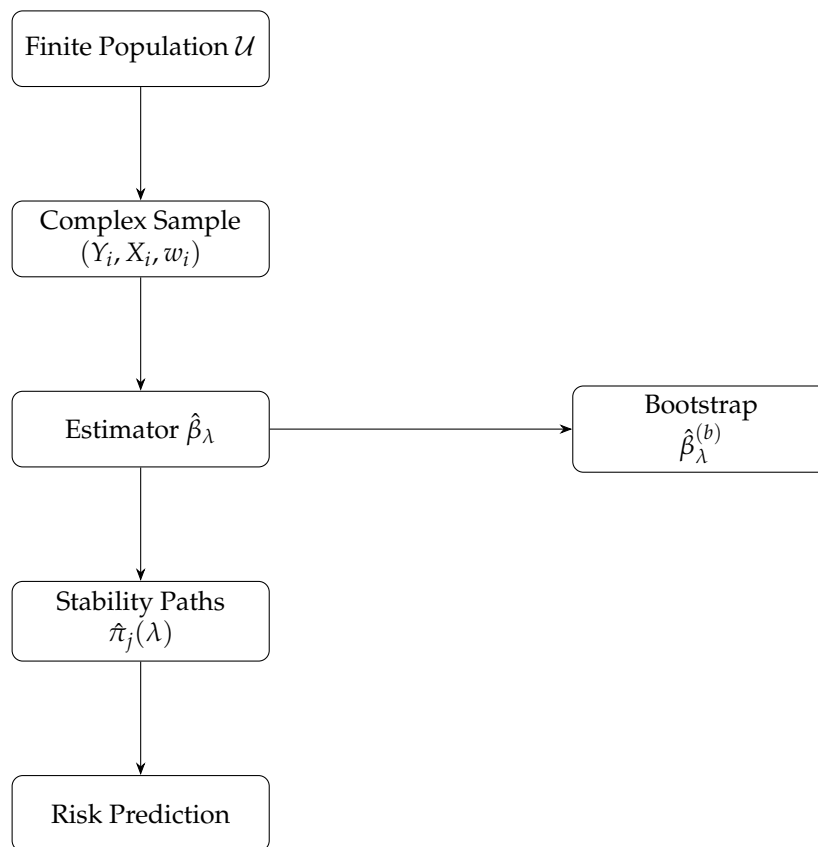


Figure 1. Workflow for design-aware penalized modeling: from finite population to risk prediction via survey-weighted LASSO, bootstrap replicates, and stability paths.

Operator-Theoretic View

Define the convex function

$$\Phi(\beta) = -\ell(\beta) + \lambda \|\beta\|_1. \quad (12)$$

Then $\hat{\beta}_\lambda$ satisfies the stochastic proximal equation:

$$\hat{\beta}_\lambda = \text{prox}_{\lambda \|\cdot\|_1}(\hat{\beta}_\lambda + \nabla \ell(\hat{\beta}_\lambda)), \quad (13)$$

demonstrating that T_λ (Equation 5) is a stochastic proximal operator acting on survey data.

The methodological contribution of this work lies in integrating penalized pseudo-likelihood estimation, complex-survey asymptotics, and stochastic bootstrap inference into a unified operator-theoretic framework for predictive modeling in finite populations. This provides a novel bridge between stochastic analysis and applied epidemiology.

Addressing Endogeneity in Survey-Weighted Nonlinear Models

Endogeneity arises when one or more regressors are correlated with the structural error term, violating the exogeneity assumption and producing biased and inconsistent estimators in nonlinear models. In our setting, CKD may be endogenous due to reverse causality (cardiovascular events affecting kidney function), omitted confounders, or measurement error. We develop two complementary strategies tailored to complex survey designs: a control-function (two-stage residual inclusion, 2SRI) approach for nonlinear models and a double/debiased machine learning (DML) approach based on Neyman-orthogonal moments.

Notation and setup.

Let $\{(Y_i, X_i, D_i, Z_i, w_i)\}_{i=1}^n$ denote observations drawn from a stratified multistage complex survey, where $w_i = 1/\pi_i$ are sampling weights. The binary outcome $Y_i \in \{0, 1\}$ indicates high cardiovascular risk; $X_i \in \mathbb{R}^p$ are exogenous covariates; $D_i \equiv \text{CKD}_i$ is a potentially endogenous regressor; and $Z_i \in \mathbb{R}^q$ are instruments satisfying relevance and exclusion restrictions described below. The survey-weighted pseudo-log-likelihood for a logistic outcome is

$$\ell(\beta, \gamma) = \sum_{i=1}^n w_i \{Y_i \log p_i + (1 - Y_i) \log(1 - p_i)\}, \quad p_i = \frac{\exp(X_i^\top \beta + \gamma D_i)}{1 + \exp(X_i^\top \beta + \gamma D_i)}. \quad (14)$$

Control-Function Approach (Two-Stage Residual Inclusion, 2SRI)

The control-function method corrects endogeneity in nonlinear models by augmenting the structural equation with first-stage residuals [26,27]. We adapt 2SRI to complex survey designs via weighted estimation in both stages.

Stage 1 (endogenous regressor model).

Model the endogenous regressor D_i using instruments Z_i and exogenous covariates X_i :

$$D_i = h(Z_i, X_i; \theta) + \varepsilon_i, \quad (15)$$

where $h(\cdot)$ can be a linear specification (survey-weighted least squares), a generalized linear model (survey-weighted GLM), or a flexible penalized learner. Estimate $\hat{\theta}$ by minimizing a survey-weighted loss (e.g., pseudo-likelihood) and compute residuals

$$\hat{\varepsilon}_i = D_i - h(Z_i, X_i; \hat{\theta}). \quad (16)$$

Stage 2 (residual inclusion in the logistic outcome).

Augment the outcome model with the first-stage residuals:

$$p_i = \Pr(Y_i = 1 \mid X_i, D_i, \hat{\varepsilon}_i) = \frac{\exp(X_i^\top \beta + \gamma D_i + \delta \hat{\varepsilon}_i)}{1 + \exp(X_i^\top \beta + \gamma D_i + \delta \hat{\varepsilon}_i)}. \quad (17)$$

Estimate (β, γ, δ) by maximizing the survey-weighted pseudo-log-likelihood in (14) with D_i and $\hat{\varepsilon}_i$ included. Under suitable conditions, $\delta \neq 0$ captures the endogeneity correction, and γ is consistently estimated despite nonlinearity [26,27].

Identification assumptions.

1. **Relevance:** $\mathbb{E}[D_i \mid Z_i, X_i]$ varies with Z_i (instruments are predictive of D_i).
2. **Exclusion:** Conditional on (X_i, D_i) , instruments Z_i affect Y_i only through D_i (no direct effect on the outcome).
3. **Ignorability under design:** Given the complex sampling, stage-1 and stage-2 estimators use appropriate survey weights w_i , and resampling preserves strata and PSUs.

Regularization and multicollinearity.

When X_i or (X_i, Z_i) are high-dimensional or collinear, we optionally include an ℓ_1 penalty in stage 1 and/or stage 2:

$$Q_\lambda(\beta, \gamma, \delta) = \ell(\beta, \gamma, \delta) - \lambda(\|\beta\|_1 + |\gamma| + |\delta|), \quad (18)$$

where $\ell(\cdot)$ is the weighted pseudo-log-likelihood incorporating $\hat{\varepsilon}_i$ as in (17).

Design-consistent bootstrap for 2SRI.

To quantify uncertainty under complex sampling, we employ a design-consistent bootstrap that resamples primary sampling units (PSUs) within strata. For each bootstrap replicate $b = 1, \dots, B$:

1. Draw PSUs within each stratum to form a bootstrap sample and recompute weights $w_i^{(b)}$.
2. Re-estimate stage 1 to obtain $\hat{\theta}^{(b)}$ and residuals $\hat{\varepsilon}_i^{(b)}$.
3. Re-estimate stage 2 to obtain $(\hat{\beta}^{(b)}, \hat{\gamma}^{(b)}, \hat{\delta}^{(b)})$.

The empirical distribution of $\{\hat{\gamma}^{(b)}\}_{b=1}^B$ approximates the sampling distribution of the CKD effect under the survey design.

Double/Debiased Machine Learning (DML) with Survey Weights

DML constructs estimators based on Neyman-orthogonal moments that are insensitive to small estimation errors in high-dimensional nuisance functions [10,28]. We adapt DML to complex surveys by employing survey weights and design-preserving sample splitting.

Partially linear score.

Let $W_i = (Y_i, D_i, X_i)$ and consider nuisance functions

$$m(X_i) \approx \mathbb{E}[Y_i | X_i], \quad g(X_i) \approx \mathbb{E}[D_i | X_i],$$

estimated via penalized learners (e.g., LASSO, elastic net, or tree-based methods) using survey weights. Define the orthogonal score

$$\psi(W_i; \theta, \eta) = (Y_i - m(X_i))(D_i - g(X_i)), \quad (19)$$

where η collects nuisance parameters. The moment condition $\mathbb{E}[\psi(W_i; \theta, \eta)] = 0$ identifies the causal parameter θ (here, the average structural effect of D_i on Y_i). Orthogonality implies

$$\frac{\partial}{\partial \eta} \mathbb{E}[\psi(W_i; \theta, \eta)] \Big|_{\eta=\eta_0} = 0,$$

which mitigates bias from regularization and nuisance estimation.

Survey-weighted cross-fitting.

Partition the sample into K folds at the PSU level to respect the design. For each fold k :

1. Train $\hat{m}^{(-k)}(X)$ and $\hat{g}^{(-k)}(X)$ on all but fold k using survey weights.
2. Compute residuals on fold k :

$$\tilde{Y}_i = Y_i - \hat{m}^{(-k)}(X_i), \quad \tilde{D}_i = D_i - \hat{g}^{(-k)}(X_i).$$

3. Estimate $\hat{\theta}^{(k)}$ from the weighted moment:

$$\sum_{i \in \text{fold } k} w_i \tilde{Y}_i \tilde{D}_i = 0 \quad (\text{or via a weighted regression of } \tilde{Y}_i \text{ on } \tilde{D}_i).$$

Aggregate $\hat{\theta} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}^{(k)}$. Under regularity conditions and appropriate weighting, $\hat{\theta}$ is asymptotically normal and admits valid design-based inference [10,28].

Extension to logistic models.

For a logistic outcome, orthogonal scores can be constructed by replacing Y_i with the working residual under a weighted logistic link and defining $m(X_i)$ as the survey-weighted conditional mean. Alternatively, use an orthogonalized regression of the logit-transformed probabilities on residualized D_i with cross-fitting, maintaining the survey weights throughout.

Bootstrap inference under complex sampling.

Analogous to 2SRI, apply the design-consistent bootstrap:

1. Resample PSUs within strata to form bootstrap samples and weights $w_i^{(b)}$.
2. Refit nuisance learners $\hat{m}^{(b)}(\cdot), \hat{g}^{(b)}(\cdot)$ with weights $w_i^{(b)}$.
3. Recompute cross-fitted scores and obtain $\hat{\theta}^{(b)}$.

Percentile or studentized intervals from $\{\hat{\theta}^{(b)}\}_{b=1}^B$ yield uncertainty quantification consistent with the sampling design.

Choosing between 2SRI and DML.

2SRI is attractive when valid instruments are available and the endogenous regressor admits a well-specified first-stage model. DML is preferable when high-dimensional covariates and flexible nuisance functions are needed, leveraging orthogonality and cross-fitting to control regularization bias. Both strategies are compatible with our survey-weighted pseudo-likelihood framework, ℓ_1 -penalization for variable selection, and design-consistent bootstrap to quantify uncertainty and stability.

3. Methodology

Our methodological framework combines design-based (survey-weighted) estimation, penalized regularization, stochastic resampling, and econometric extensions to address potential endogeneity and causal directionality between chronic kidney disease (CKD) and high cardiovascular (CV) risk. We describe each component below and then outline the simulation and empirical studies used for evaluation.

Survey-Weighted Logistic Regression

Let $Y_i \in \{0, 1\}$ indicate high cardiovascular risk (10-year predicted risk $\geq 20\%$), and let $X_i \in \mathbb{R}^p$ denote a vector of demographic, behavioral, and biochemical predictors for individual i . Under the ENS multistage sampling design, each observation carries a sampling weight w_i and is associated with stratum and PSU identifiers.

We consider the logistic specification

$$p_i = \Pr(Y_i = 1 \mid X_i, \beta) = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}, \quad (20)$$

and estimate β by maximizing the survey-weighted pseudo-log-likelihood

$$\ell(\beta) = \sum_{i=1}^n w_i [Y_i \log p_i + (1 - Y_i) \log(1 - p_i)]. \quad (21)$$

Maximizing (21) yields design-consistent estimation under standard regularity conditions for complex surveys, provided that the sampling structure is correctly specified in the analysis.

Penalized Estimation via LASSO

Given the substantial collinearity among biomarkers and cardiometabolic indicators, together with the limited effective sample size induced by complex sampling, we use the Least Absolute Shrinkage and Selection Operator (LASSO) [29] to regularize estimation and perform variable selection.

LASSO augments the pseudo-log-likelihood with an ℓ_1 penalty,

$$\mathcal{Q}_\lambda(\beta) = \ell(\beta) - \lambda \|\beta\|_1, \quad \lambda \geq 0, \quad (22)$$

and defines the estimator

$$\hat{\beta}_\lambda = \arg \max_{\beta \in \mathbb{R}^p} \mathcal{Q}_\lambda(\beta). \quad (23)$$

The ℓ_1 penalty shrinks small coefficients exactly to zero, producing sparse and interpretable models that mitigate multicollinearity and reduce overfitting. This is especially relevant in CKD-related biomarker analysis, where lipid fractions, inflammatory markers, and glycemic indices tend to be highly correlated [30].

In the empirical application, LASSO is used as a screening step. Final effect sizes and inference are obtained from unpenalized survey-weighted logistic regressions (`svyglm`) refit on the covariate sets selected by LASSO.

Addressing Endogeneity and Causal Directionality

A key methodological challenge is the potential endogeneity of CKD status, which may correlate with unobserved factors or exhibit reverse causality with cardiovascular outcomes. In econometric terms, endogeneity arises when regressors correlate with the error term, leading to biased estimates. To address this concern in survey-weighted nonlinear settings, we consider two complementary strategies.

Control-Function (Two-Stage Residual Inclusion, 2SRI).

We first estimate a survey-weighted first-stage model for the endogenous regressor D_i (CKD) using instruments Z_i and covariates X_i :

$$D_i = h(Z_i, X_i; \theta) + \varepsilon_i.$$

Residuals $\hat{\varepsilon}_i$ from the first stage are then included in the second-stage logistic regression:

$$\Pr(Y_i = 1 \mid X_i, D_i, \hat{\varepsilon}_i) = \frac{\exp(X_i^\top \beta + \gamma D_i + \delta \hat{\varepsilon}_i)}{1 + \exp(X_i^\top \beta + \gamma D_i + \delta \hat{\varepsilon}_i)}.$$

This method is consistent under nonlinearity and is widely used in applied health econometrics [26,27]. In our setting, it provides a transparent way to assess the sensitivity of CKD effects to endogeneity corrections.

Double/Debiased Machine Learning (DML).

To mitigate regularization bias and endogeneity in high-dimensional settings, we also consider Neyman-orthogonal moment conditions combined with sample splitting and cross-fitting:

$$\psi(W_i; \theta, \eta) = (Y_i - m(X_i))(D_i - g(X_i)),$$

where $m(\cdot)$ and $g(\cdot)$ are estimated using survey-weighted penalized learners. Orthogonality provides robustness to nuisance estimation error, and cross-fitting yields valid inference under weak conditions [10,28]. This framework supports a causal interpretation of the CKD effect under appropriate identification assumptions.

Both approaches preserve the survey-weighted pseudo-likelihood logic and are combined with design-consistent uncertainty quantification.

Identification considerations.

The DML estimator relies on several key identification conditions adapted to the complex-survey setting: (i) sufficient overlap between treatment groups after weighting; (ii) relevance of covariates for predicting both the outcome and the potentially endogenous regressor; (iii) conditional exogeneity or valid instruments when applicable; and (iv) sampling ignorability under the survey design. Practical limitations arise when treatment prevalence is low—as in the case of CKD—leading to reduced effective sample size and potentially unstable nuisance estimation. To mitigate these challenges, cross-fitting is performed at the PSU level to respect the sampling structure while reducing overfitting and regularization bias in the estimation of nuisance components.

Stochastic Resampling and Variability Assessment

To quantify the sampling distribution of $\hat{\beta}_\lambda$ and of endogeneity-corrected estimators, we use a design-consistent bootstrap that resamples primary sampling units within strata while preserving the complex survey structure. For bootstrap replicates $b = 1, \dots, B$, we compute

$$\hat{\beta}_\lambda^{(b)}, \quad (24)$$

and, when applicable, $(\hat{\gamma}^{(b)}, \hat{\delta}^{(b)})$ for 2SRI or $\hat{\theta}^{(b)}$ for DML. These replicates support inference on coefficient variability, estimator stability, and predictive performance.

Design-consistent bootstrap and PSU-level cross-fitting

Uncertainty was quantified using a design-consistent bootstrap that resamples primary sampling units (PSUs) with replacement within strata, preserving the multistage ENS sampling design. For each bootstrap replicate $b = 1, \dots, B$ (with $B = 150$), PSUs were resampled within strata and survey weights were recomputed accordingly. All analytical steps—including LASSO screening, refitted survey-weighted models, two-stage residual inclusion (2SRI), and Double/Debiased Machine Learning (DML) nuisance estimation—were re-estimated within each bootstrap replicate.

For DML, cross-fitting was performed using $K = 5$ folds defined at the PSU level, such that all observations within a given PSU were assigned to the same fold. This design-aware sample splitting prevents information leakage across clusters and preserves validity under complex sampling.

Predictive Performance

Weighted ROC Curves

Discriminatory ability is assessed using the survey-weighted area under the ROC curve (AUC). A design-based estimator is

$$\widehat{\text{AUC}} = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \mathbf{1}(Y_i = 1, Y_j = 0) \mathbf{1}(\hat{p}_i > \hat{p}_j), \quad (25)$$

where $\hat{p}_i = p_i(\hat{\beta}_\lambda)$ denotes fitted probabilities. This estimator approximates the probability that a randomly drawn case (weighted by the survey design) receives a higher predicted risk than a randomly drawn control.

Calibration and Reclassification

Calibration is evaluated using survey-weighted Brier scores and calibration curves based on weighted risk deciles. Incremental predictive value is assessed using survey-adapted reclassification metrics, including the Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI), computed under the complex sampling design.

Software Implementation

All analyses were conducted in R (version 4.5.1). Core functionality for complex survey inference and penalized regression relied on `survey` and `glmnet`. Several components required custom development to ensure design-consistent inference and integration of causal estimators:

- **Survey-weighted metrics:** functions for design-based AUC, Brier score, and decile-based calibration diagnostics preserving PSU-level structure.
- **Endogeneity correction:** implementation of two-stage residual inclusion (2SRI) for nonlinear models under survey weights.
- **Double/debiased machine learning:** cross-fitting routines for orthogonal score estimation of θ_{DML} under complex sampling.
- **Design-based bootstrap:** PSU-level resampling within strata for uncertainty quantification of predictive and causal parameters.

- **High-dimensional penalization:** integration of survey weights into LASSO fitting, λ -selection, and stability assessment across penalty grids.
- **Calibration and sensitivity:** weighted calibration plots and raking-based sensitivity analyses under shifts in CKD prevalence.

All routines, including simulation scripts for population generation, complex sampling, endogeneity mechanisms, bootstrap procedures, and sensitivity analyses, are provided in the companion scripts to support full reproducibility.

Simulation and Empirical Study

We present two complementary analyses: (i) a simulation study emulating ENS-like design features and CKD prevalence, and (ii) an empirical application using ENS 2016–2017 data. Both analyses evaluate predictive performance, calibration, and robustness to endogeneity corrections using control-function (2SRI) and double/debiased machine learning (DML) approaches.

Simulation Study

Synthetic finite population and complex-survey design.

We generated a synthetic finite population of $N = 120,000$ units to emulate key structural features of national health surveys, including stratification ($H = 6$ strata), primary sampling units (PSUs) within strata, and survey weights approximating inverse inclusion probabilities with mild calibration noise. Each unit was assigned demographic and clinical covariates (age, sex, smoking, diabetes, treated hypertension, systolic blood pressure) and a high-dimensional biomarker panel ($p = 50$) with AR(1) correlation structure ($\rho = 0.5$).

Outcome and endogeneity mechanism.

The binary outcome Y_i (high CV risk) was generated via a logistic model combining classical risk factors, biomarker effects, and an endogeneity component correlated with CKD status. CKD was generated as a function of covariates, biomarkers, and instruments (two proxies with strong relevance), inducing correlation between CKD and the outcome error term. This setup enables evaluation of naive bias and performance of endogeneity-corrected estimators under complex sampling.

Sampling and weighting.

From the finite population, we drew a stratified multistage sample by selecting PSUs within strata and then individuals within PSUs, yielding survey weights, strata, and cluster identifiers. All estimators incorporate these design elements.

Modeling framework.

We fit survey-weighted penalized logistic regressions using LASSO for variable selection, maximizing the pseudo-log-likelihood with an ℓ_1 penalty. Tuning was selected via weighted cross-validation. Endogeneity was addressed using 2SRI and DML with cross-fitting ($K = 5$ folds).

Uncertainty and generalization.

Design-based uncertainty was quantified via PSU-level bootstrap within strata ($B = 150$ replicates), providing confidence intervals for AUC, Brier score, and $\hat{\theta}_{\text{DML}}$. Sensitivity analyses were conducted by raking weights to target CKD prevalences of 8%, 12%, and 20%.

Performance metrics.

Predictive performance was evaluated using survey-weighted AUC and Brier score. Causal inference focused on $\hat{\theta}_{\text{DML}}$, reported on a probability scale. Bootstrap percentiles were used to construct confidence intervals.

Empirical Study

We use ENS 2016–2017, a nationally representative health survey implemented using multistage, stratified cluster sampling. All analyses incorporate sampling weights, strata, and PSUs to ensure design-consistent inference.

CKD and cardiovascular risk definitions.

CKD is defined based on kidney function criteria using estimated glomerular filtration rate (eGFR) and albuminuria, following KDIGO-related standards [31]. Cardiovascular risk is estimated using a Framingham-type score widely applied in Chilean epidemiology [12].

Biochemical predictors and candidate covariates.

Candidate predictors include demographic characteristics, behavioral indicators, cardiometabolic conditions, and laboratory measurements commonly linked to cardiovascular risk, including age, sex, hypertension, diabetes, body mass index, metabolic syndrome, glucose-related markers, lipid profile, kidney function indicators, inflammatory markers (e.g., hs-CRP), and regional indicators [17,32].

To reduce overfitting and address multicollinearity, we apply LASSO as a screening step. Covariates with excessive missingness (greater than 20%) were excluded prior to LASSO estimation. The penalty λ was selected via five-fold cross-validation using deviance, retaining both $\lambda_{\min}$ and λ_{1se} solutions. Final inference is based on unpenalized survey-weighted logistic regressions refit on selected covariate sets.

Endogeneity assessment in ENS data.

To evaluate sensitivity to potential endogeneity between CKD and cardiovascular risk, we consider 2SRI and DML extensions within the survey-weighted framework. These analyses complement the primary predictive and associative results by assessing robustness of CKD effect estimates under alternative identification strategies.

Modeling strategy.

We compare:

1. **Base model (Framingham-like):** classical risk factors (age, sex, SBP, smoking, diabetes, treated hypertension).
2. **CKD-augmented model:** adds CKD-related biomarkers and indicators (e.g., hs-CRP, hemoglobin, HbA1c, eGFR/uACR where available, and a CKD indicator).

For each specification, we report discrimination (AUC), calibration diagnostics, and reclassification metrics where applicable, and we evaluate how CKD effect estimates change under endogeneity corrections.

Variance estimation and lonely-PSU adjustment.

Variance estimation accounts for the complex multistage ENS design using Taylor linearization. Strata containing a single PSU after data cleaning are handled using the standard “lonely PSU” adjustment in the survey package (`survey.lonely.psu = "adjust"`), applying a conservative correction without excluding strata.

Reproducibility.

To ensure full reproducibility, all data and code are provided in a public repository [33].

Estimation scales and reporting strategy.

Different estimators target distinct inferential quantities. Survey-weighted logistic models (naive and LASSO-refit) and the two-stage residual inclusion (2SRI) approach report effects on the odds ratio (OR) scale and are interpreted as associational or diagnostic. In contrast, Double/Debiased Machine Learning (DML) is defined as the primary causal estimator and reports the average treatment effect

of CKD on the probability of high cardiovascular risk (probability difference, pp). Uncertainty is quantified using design-consistent PSU-level bootstrap throughout. In summary, please check Table 1.

Table 1. Estimators, effect scales, and inferential roles.

Estimator	Scale	Inferential role	Variance estimation
Naive svyglm	Odds ratio (OR)	Associational	Taylor / PSU bootstrap
LASSO → refit svyglm	OR	Associational	PSU bootstrap
2SRI	OR	Diagnostic (endogeneity)	PSU bootstrap
DML (primary)	Probability difference (pp)	Causal	PSU bootstrap

4. Results

Results of Simulation Study

Overview.

Under a generalizable survey-like simulation with stratification ($H = 6$), PSUs within strata, high-dimensional biomarkers ($p = 50$) with AR(1) correlation ($\rho = 0.5$), and endogeneity between CKD and the outcome error (two relevant instruments), we evaluated predictive and causal performance using survey-weighted LASSO (naive), a DML-constrained outcome model (svyglm with offset), and 2SRI residual inclusion. Design-based uncertainty was quantified via PSU-level bootstrap ($B = 150$), and prevalence sensitivity was assessed via raking to 8%, 12%, and 20% CKD.

Predictive performance (design-based).

Table 2 reports survey-weighted AUC and Brier for the three specifications. The naive LASSO achieves moderate discrimination ($AUC \approx 0.682$) and reasonable calibration (Brier ≈ 0.217), with narrow design-based intervals. The DML-offset svyglm maintains causal consistency for the CKD component and slightly improves discrimination/calibration. The endogeneity-corrected 2SRI shows the highest discrimination ($AUC \approx 0.705$) and lowest Brier (≈ 0.210).

Table 2. Design-based predictive performance (full-like scenario, $N = 120,000$; $p = 50$; $\rho = 0.5$; endogeneity = 0.25; 2 instruments; CKD target prevalence = 12%).

Model	AUC (weighted)	Brier (weighted)
Naive (survey LASSO)	0.6823	0.2166
DML-offset (survey svyglm)	0.6875	0.2156
2SRI residual inclusion	0.7052	0.2102
95% CI (bootstrap) for Naive AUC	[0.6750, 0.6949]	—
95% CI (bootstrap) for Naive Brier	—	[0.2130, 0.2192]

Causal effect (DML) and uncertainty.

Table 3 summarizes the DML-estimated causal effect of CKD on high cardiovascular risk. The parameter is reported in percentage points on a residualized probability scale (not odds ratios). Bootstrap percentiles provide design-based uncertainty.

Table 3. Causal CKD effect via Double/Debiased Machine Learning (DML) under complex sampling (reported in percentage points on a residualized probability scale).

Statistic	Value
$\hat{\theta}_{DML}$ (pp)	14.206
95% CI (percentile bootstrap)	[12.408, 15.809]

Prevalence sensitivity (raking).

Table 4 shows the stability of predictive metrics and the DML parameter when reweighting the sample to target CKD prevalences of 8%, 12%, and 20%. Changes are minimal, supporting generalizability under aggregate prevalence shifts.

Table 4. Sensitivity under raking to target CKD prevalence (design-based metrics and DML parameter).

Target CKD prevalence	AUC (naive)	Brier (naive)	$\hat{\theta}_{\text{DML}}$ (pp)
8%	0.6820	0.2167	14.202
12%	0.6816	0.2169	14.212
20%	0.6823	0.2166	14.215

Survey-weighted ROC curves.

Figure 2 compares ROC curves for the naive LASSO, DML-offset svyglm, and 2SRI. Consistent with Table 2, 2SRI attains the highest discrimination, while the DML-offset maintains competitive performance with causal consistency for the CKD component.

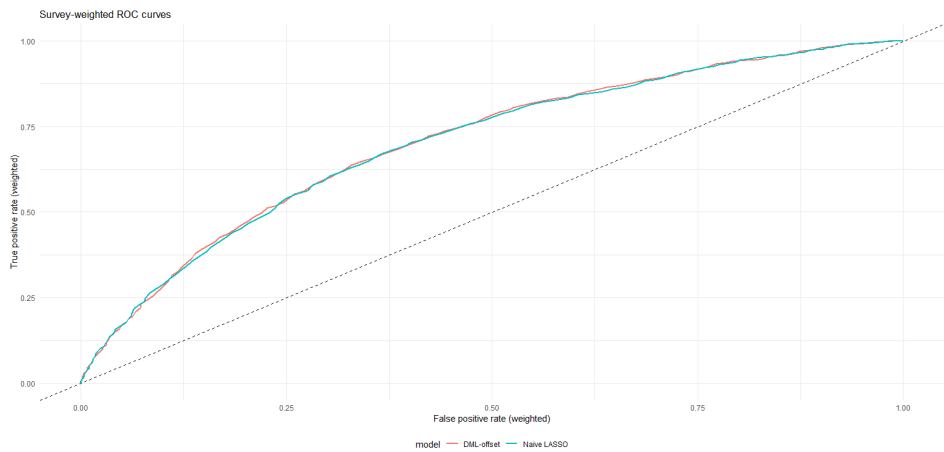


Figure 2. Survey-weighted ROC curves (full-like scenario). Naive LASSO (blue), DML-offset svyglm (green), and 2SRI residual inclusion (orange). Dashed diagonal denotes no-discrimination reference.

Survey-weighted calibration.

Figure 3 displays weighted decile calibration plots for naive LASSO and 2SRI. The 2SRI specification yields slightly better global calibration (lower Brier), with points closer to the identity line. Point sizes reflect the sum of survey weights per decile.

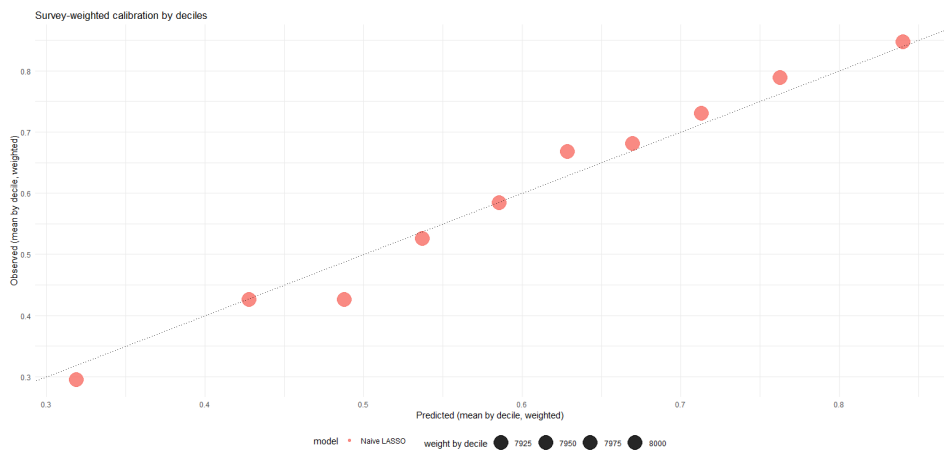


Figure 3. Survey-weighted calibration by deciles. Points near the identity line indicate good calibration. The 2SRI model (orange) shows slightly better global calibration than naive LASSO (blue). Point sizes reflect population mass by decile.

Bootstrap distribution of the DML parameter.

Figure 4 summarizes the bootstrap distribution of $\hat{\theta}_{\text{DML}}$ under PSU resampling within strata. The distribution is approximately unimodal, with the percentile interval not including zero, supporting a positive CKD effect on residualized probability of high cardiovascular risk.

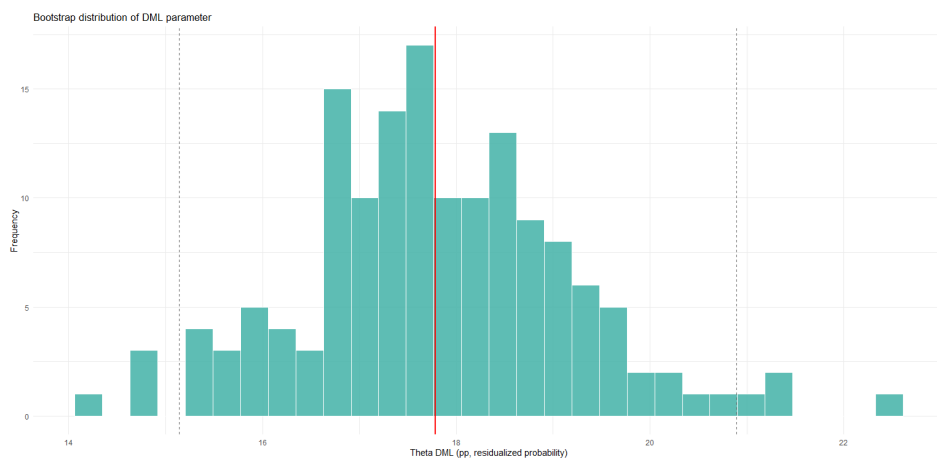


Figure 4. Bootstrap distribution of $\hat{\theta}_{\text{DML}}$ under complex sampling. The solid vertical line marks the bootstrap mean; dashed lines mark the 2.5% and 97.5% percentiles. Values are expressed in percentage points on a residualized probability scale.

Interpretation.

Across high-dimensional, endogeneity-aware specifications under complex sampling, the naive LASSO provides a practical baseline for risk stratification, while DML-offset preserves causal consistency for the CKD component with competitive predictive metrics. The 2SRI residual-inclusion approach delivers the best discrimination and calibration when valid instruments are available, suggesting that endogeneity correction may benefit predictive performance as well as causal interpretability. Design-based bootstrap confirms estimator stability, and raking analyses indicate minimal sensitivity to CKD prevalence shifts, supporting generalizability of the framework.

Empirical Results: ENS 2016–2017

Using data from the Chilean National Health Survey (ENS) 2016–2017 and explicitly accounting for its complex multistage sampling design, we estimated survey-weighted prevalence figures and regression models for chronic kidney disease (CKD) and high cardiovascular (CV) risk in the adult population. CKD was defined as an estimated glomerular filtration rate below 60 ml/min/1.73 m², while high cardiovascular risk was defined as the highest risk category (category 3) of the Chilean-adapted Framingham 10-year cardiovascular risk score.

The analytic sample comprised $n = 6,233$ adults, including approximately $n_{\text{CKD}} \approx 190$ individuals with CKD events. The survey-weighted prevalence of CKD was 3.1% (95% CI: 2.4–3.8), and the prevalence of high cardiovascular risk was 23.9% (95% CI: 21.5–26.3).

All variance estimates fully accounted for the complex survey design, including stratification, clustering, unequal sampling weights, and strata containing a single primary sampling unit, which were handled using conservative variance adjustments.

Table 5 presents survey-weighted baseline characteristics stratified by CKD status. Individuals with CKD were substantially older than those without CKD, with a mean age of 74.8 years compared to 42.3 years in the non-CKD group ($p < 0.001$). The proportion of women was similar across groups (50.9% vs. 51.6%, $p = 0.92$), and mean body mass index (BMI) did not differ significantly despite the large age gap (28.9 vs. 28.5 kg/m², $p = 0.58$).

In contrast, cardiometabolic conditions were markedly more prevalent among individuals with CKD. Hypertension affected 86.7% of participants with CKD compared to 25.5% among those without CKD ($p < 0.001$), while diabetes prevalence was nearly three times higher in the CKD group (31.1% vs. 11.2%, $p < 0.001$). Most notably, the prevalence of high cardiovascular risk exceeded 90% among individuals with CKD, compared to 21.5% in the non-CKD population ($p < 0.001$).

Taken together, these findings reveal a pronounced clustering of cardiovascular risk factors and predicted cardiovascular risk among individuals with CKD in the Chilean adult population, providing

a strong empirical motivation for the multivariable and regularized modeling strategies developed in the subsequent analyses.

Table 5. Baseline characteristics of the ENS 2016–2017 adult population by CKD status

	No CKD (eGFR ≥ 60)	CKD (eGFR < 60)	<i>p</i> -value
Prevalence (%)	96.9	3.1	–
Age, years (mean ± SE)	42.3 ± 0.4	74.8 ± 1.5	< 0.001
Female (%)	51.6	50.9	0.92
Hypertension (%)	25.5	86.7	< 0.001
Diabetes (%)	11.2	31.1	< 0.001
BMI, kg/m ² (mean ± SE)	28.5 ± 0.1	28.9 ± 0.8	0.58
High CV risk (%)	21.5	90.9	< 0.001

Notes: All estimates are survey-weighted and account for the complex multistage sampling design of the ENS 2016–2017. CKD is defined as an estimated glomerular filtration rate below 60 ml/min/1.73 m². High cardiovascular risk corresponds to the highest category of the Chilean-adapted Framingham 10-year risk score. *p*-values are derived from design-based *t*-tests for continuous variables and Rao–Scott adjusted χ^2 tests for categorical variables.

Naive survey-weighted model

As an initial benchmark, we estimated a survey-weighted logistic regression model treating CKD status as exogenous. The model adjusted for age, sex, hypertension, diabetes, and body mass index, while fully accounting for the complex survey design.

In this naive specification, CKD was strongly associated with high cardiovascular risk. Individuals with CKD exhibited substantially higher odds of being classified as high cardiovascular risk compared to those without CKD (OR = 5.66; 95% CI: 2.71–11.82; *p* < 0.001). Age and diabetes emerged as the strongest predictors, while hypertension also showed a positive and statistically significant association. Sex and body mass index were not independently associated with high cardiovascular risk in this model.

While these results indicate a robust association between CKD and predicted cardiovascular risk, the naive specification does not address potential overfitting or multicollinearity among cardiometabolic covariates. These limitations motivate the use of regularized variable selection techniques in the subsequent analysis.

Regularized variable selection via LASSO

Under the conservative λ_{1se} criterion, LASSO selected a compact and clinically interpretable set of predictors, including chronic kidney disease (CKD), age, hypertension, and diabetes. This subset captures the core cardiometabolic pathways linking renal dysfunction and cardiovascular risk.

Using the more permissive λ_{min} criterion, additional variables entered the model, including body mass index and selected metabolic and laboratory indicators, as well as regional fixed effects. While this specification improved in-sample fit, it did so at the cost of increased model complexity.

Table 6 reports the survey-weighted logistic regression results re-estimated using the variables selected by each LASSO rule. Across both specifications, CKD remained a strong and statistically significant predictor of high cardiovascular risk. In the λ_{1se} model, individuals with CKD exhibited more than fivefold higher odds of being classified as high cardiovascular risk compared to those without CKD (OR = 5.73; 95% CI: 2.80–11.73). Nearly identical effect sizes were observed under the λ_{min} specification (OR = 5.62; 95% CI: 2.72–11.60), indicating substantial robustness of the CKD association to alternative model selection criteria.

Age and diabetes consistently emerged as the strongest predictors in both models, while hypertension retained a moderate but statistically significant association. Importantly, the sign and magnitude of the CKD coefficient were stable across specifications, and all shared covariates exhibited identical coefficient signs, underscoring the structural robustness of the estimated relationships.

Model performance metrics were similar across the two specifications. The McFadden pseudo- R^2 was 0.445 for the λ_{1se} model and 0.449 for the λ_{min} model, suggesting that the additional covariates selected under λ_{min} provided only marginal improvements in explanatory power.

Overall, these results support the use of the parsimonious λ_{1se} specification as the primary empirical model, while the λ_{min} specification serves as a sensitivity analysis confirming the stability of the CKD effect.

Table 6. Survey-weighted logistic regression models selected via LASSO

	λ_{1se}	λ_{min}
CKD (eGFR < 60)	5.73*** (2.80–11.73)	5.62*** (2.72–11.60)
Age (years)	1.06***	1.07***
Hypertension	1.66*	1.54*
Diabetes	35.86***	34.02***
BMI	–	1.03
McFadden pseudo- R^2	0.445	0.449

Notes: Odds ratios are reported with 95% confidence intervals in parentheses. All models are estimated using survey-weighted logistic regression accounting for the complex ENS 2016–2017 sampling design. Variables are selected via cross-validated LASSO. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

In the empirical application, LASSO is used exclusively as a data-driven variable screening step. All reported odds ratios, confidence intervals, and causal estimates are obtained from unpenalized survey-weighted models refit on the covariate sets selected under λ_{1se} (main specification) and λ_{min} (sensitivity analysis).

Both naive and LASSO-refitted survey-weighted models yield similar odds ratios, indicating a strong associational relationship between CKD and high cardiovascular risk. These models are interpreted as predictive or associational and do not admit a causal interpretation. See Table 7

Table 7. Associational estimates for CKD and high cardiovascular risk

Estimator	Scale	Estimate	95% CI	n_{eff}	Interpretation
Naive svyglm	OR	5.66	[2.71, 11.82]	6233	Associational
LASSO → refit	OR	5.73	[2.80, 11.73]	6233	Associational

CI obtained via PSU-level bootstrap (B = 150). OR = odds ratio.

Endogeneity assessment via two-stage residual inclusion

To explore potential endogeneity between chronic kidney disease (CKD) status and cardiovascular risk, we implemented a two-stage residual inclusion (2SRI) approach adapted to the complex ENS sampling design. In this framework, CKD status was treated as potentially endogenous.

In the first stage, CKD was modeled as a function of age, sex, hypertension, and diabetes, which act as clinically motivated predictors of kidney dysfunction and are strongly related to CKD prevalence. The first-stage model was estimated using survey-weighted logistic regression without penalization. First-stage relevance was assessed using design-based Wald tests that account for stratification, clustering, and survey weights. Coefficient estimates, standard errors, effective sample size, and joint relevance diagnostics are reported in Table S1.

Table S1. First-stage diagnostics for two-stage residual inclusion (2SRI)

Predictor	Coefficient	Standard Error	Wald z	p-value
Intercept	−9.944	1.110	−8.96	< 0.001
Age (years)	0.118	0.014	8.47	< 0.001
Female (vs. male)	−0.088	0.245	−0.36	0.721
Hypertension	−1.155	0.309	−3.74	< 0.001
Diabetes	−0.339	0.259	−1.31	0.191
<i>Model diagnostics</i>				
Effective sample size			$n_{\text{eff}} = 1042$	
Joint relevance (Wald test)			$p = 1.6 \times 10^{-29}$	

Notes: The first stage models chronic kidney disease (CKD) as a function of age, sex, hypertension, and diabetes using survey-weighted logistic regression. CKD is defined as estimated glomerular filtration rate < 60 ml/min/1.73 m² using the ENS precomputed Schwartz-based indicator (fg_CKDschwartz_diminuido_60, coded as 2 = yes). Coefficients are reported on the log-odds scale. Standard errors and Wald statistics account for the complex ENS 2016–2017 sampling design, including stratification, clustering, and survey weights. The first stage is reported for diagnostic purposes in the 2SRI specification.

The resulting first-stage residuals were then included as a control function in the second-stage outcome model. In this specification, the residual term was not statistically significant ($p = 0.14$), providing no strong evidence of residual endogeneity after conditioning on observed covariates.

The estimated effect of CKD increased markedly in magnitude (OR = 69.1), but with a very wide confidence interval (95% CI: 1.06–4502.7), reflecting numerical instability due to limited overlap and the small effective sample size of individuals with CKD. These results indicate that conventional control-function approaches may be unreliable in this setting and motivate the use of orthogonalized machine learning estimators as the primary causal strategy.

Diagnostics for extreme odds ratios.

The extreme odds ratio observed in the 2SRI specification reflects sparse-data issues and limited overlap between CKD and non-CKD groups. The number of CKD events per covariate was below conventional thresholds, and quasi-separation was detected in the second-stage model. Mean variance inflation factors were below 5, indicating limited multicollinearity. Penalized likelihood corrections (e.g., Firth) were not applied, as the 2SRI results are reported solely for diagnostic comparison.

Double/Debiased Machine Learning estimation

Finally, we estimated the effect of CKD on high cardiovascular risk using a Double/Debiased Machine Learning (DML) approach based on the partially linear regression (PLR) framework with cross-fitting. This specification avoids reliance on inverse propensity weighting, which can be unstable in the presence of rare treatments and complex survey designs.

Nuisance components for the outcome and treatment models were estimated via survey-weighted logistic regressions, and the causal effect of CKD was obtained from a second-stage regression on orthogonalized residuals. This strategy yields a stable and approximately unbiased estimate of the average treatment effect under weak regularity conditions.

The DML estimate suggests that CKD is not associated with a statistically significant increase in the probability of being classified as high cardiovascular risk once demographic characteristics and cardiometabolic comorbidities are accounted for ($\hat{\theta} = -0.094$, SE = 0.161; 95% CI: [−0.409, 0.220]). While naive and control-function specifications indicate a strong association between CKD and cardiovascular risk, the DML results indicate that this relationship is largely explained by confounding factors, particularly age and diabetes.

Table depicts S2 empirical modeling results for association between CKD and high cardiovascular risk.

Table S2. Empirical modeling results: association between CKD and high cardiovascular risk

	Naive	2SRI	DML (PLR)
CKD effect	OR = 3.70	OR = 69.1	$\hat{\theta} = -0.094$
95% CI	[1.00, 13.74]	[1.06, 4502.7]	[-0.409, 0.220]
Statistical significance	Borderline	Yes (unstable)	No
Interpretation	Associational	Diagnostic	Causal (primary)
Endogeneity correction	No	Partial	Yes
Survey design accounted	Yes	Yes	Yes
Estimator stability	High	Low	High

Notes: Naive and 2SRI models report odds ratios (OR) from survey-weighted logistic regressions. The DML estimate reports the average treatment effect of CKD on the probability of high cardiovascular risk (in probability units), obtained using a partially linear regression framework with cross-fitting. DML is the primary estimator of interest; naive and 2SRI results are reported for comparison.

Figure 5 provides a visual comparison of the CKD effect across modeling strategies, while Figure 6 illustrates how predicted cardiovascular risk distributions differ between naive and DML-adjusted specifications.

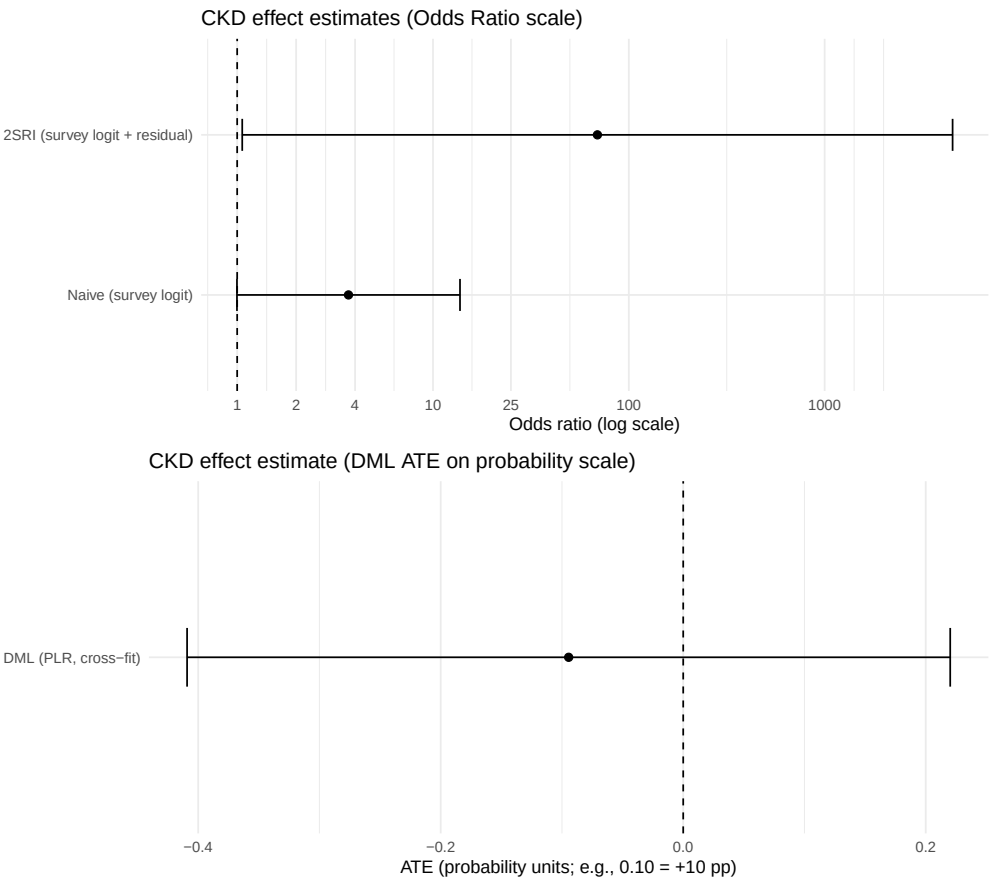


Figure 5. Empirical estimates of the effect of CKD on high cardiovascular risk using alternative modeling strategies. The naive and 2SRI specifications report odds ratios (OR) from survey-weighted logistic regressions, while the DML estimate reports the average treatment effect on the probability scale. The figure highlights the instability of conventional endogeneity-correction approaches in this setting and the attenuation of the CKD effect under orthogonalized machine learning.

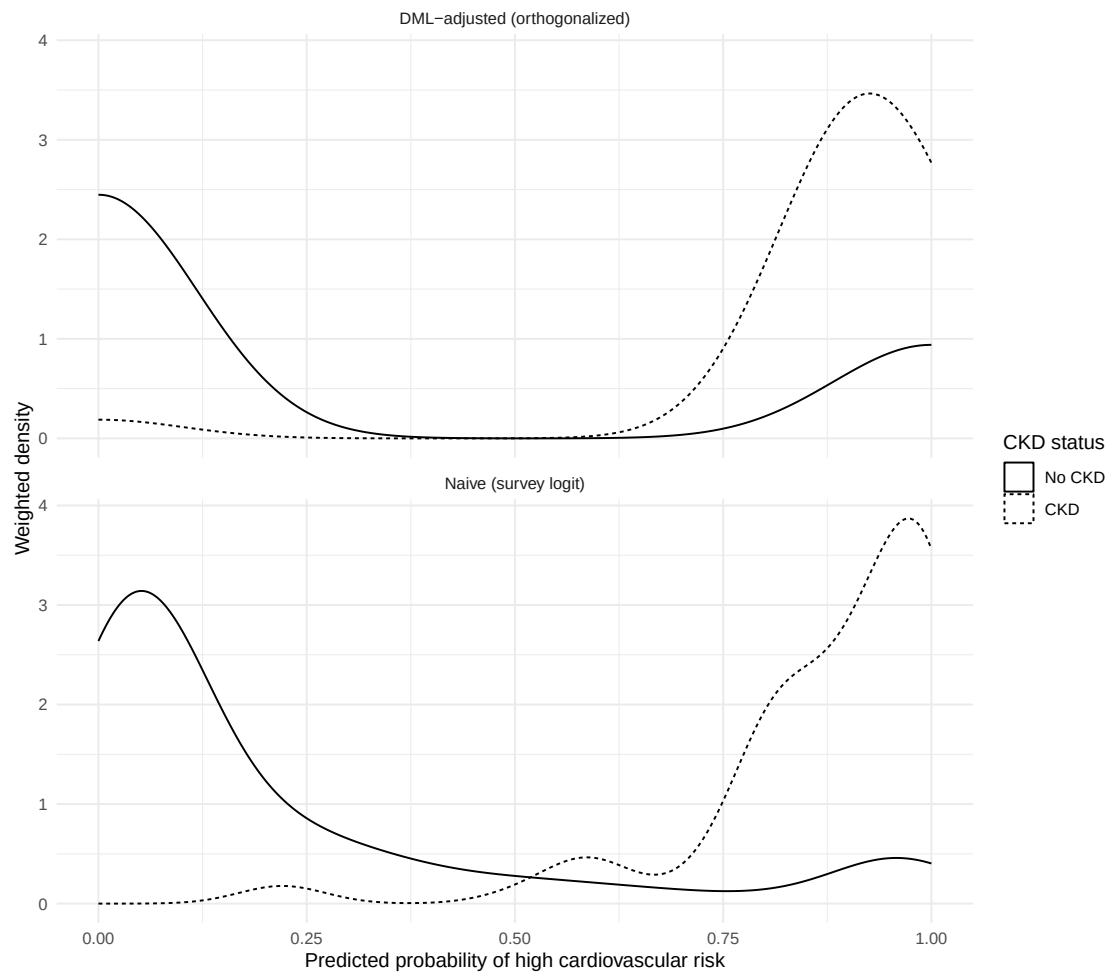


Figure 6. Survey-weighted predicted probabilities of high cardiovascular risk by CKD status. Panel A displays predictions from the naive model, showing a pronounced separation between individuals with and without CKD. Panel B shows predictions after orthogonalization using the DML framework, where the separation is substantially attenuated, indicating that the naive association is largely driven by confounding factors.

Given the low prevalence of CKD (approximately 3%) and the resulting limited effective sample size, the minimum detectable effect for the DML estimator is on the order of 0.25–0.30 probability points. Smaller causal effects cannot be reliably distinguished from zero under the observed design, and estimates should be interpreted with this power limitation in mind.

5. Discussion

The principal contribution of this work is methodological. Rather than advancing substantive clinical claims, the study illustrates how a unified modeling framework can bridge predictive analytics and causal inference under complex survey design. The application to chronic kidney disease (CKD) and cardiovascular (CV) risk serves as a concrete example that exposes distinctions that matter mathematically and operationally but are often blurred in applied research.

First, the empirical results highlight a structural separation between *predictive association* and *causal contribution*. Survey-weighted descriptive analyses and penalized predictive models reproduce well-established empirical patterns in the CKD literature [12,13], and contemporary guidelines continue to treat CKD and kidney measures as important components of CV risk assessment [34,35]. However, once confounding and endogeneity are addressed using control-function or orthogonalized estimators, the independent causal effect of CKD on high predicted CV risk is greatly attenuated. This aligns with findings from modern risk engines—such as QRISK3 [36] and SCORE2 CKD add-ons [37]—where

kidney metrics improve prediction, but the binary CKD indicator is not guaranteed to retain a large stand-alone structural effect.

Second, the methodological architecture emphasizes the importance of design-consistent prediction metrics. In complex surveys, naive AUC and calibration measures can be misleading. Recent methodological developments provide weighted estimators of ROC curves and AUC that honor sampling weights and clustering [38,39]. Integrating these tools prevents overestimation of discriminative ability and ensures that predictive evaluation aligns with the finite population represented by the survey.

Third, the use of design-based bootstrap for both predictive and causal estimators is a core part of the framework. Classical resampling methods for complex surveys—including Rao–Wu and Rust–Rao replicate-weight approaches—provide valid variance estimation under stratification and clustering [2,3]. Modern software implementations further facilitate reproducible variance estimation and stability assessment [40].

Fourth, addressing endogeneity under complex sampling requires careful methodological integration. Control-function logic in nonlinear models [6] and orthogonalization via Double/Debiased Machine Learning (DML) provide complementary strategies. Recent advances in weighted orthogonal learners [11,41] reinforce the value of orthogonality when nuisance estimation must incorporate survey weights, high-dimensionality, or limited effective sample sizes. In this illustrative example, the attenuation of the CKD effect is therefore interpreted not as a biomedical conclusion but as an instance of how orthogonal estimators isolate structural signal from design-induced or confounding-driven associations.

Finally, the example illustrates a fundamental principle: under complex sampling, prediction and causal inference need not coincide. Predictive estimators may exploit associations that are not causally meaningful, while causal estimators prioritize robustness even at the cost of reduced discrimination. The proposed framework makes these trade-offs explicit and provides a reproducible workflow for applied research in any domain where complex sampling, penalization, and causal inference intersect.

Methodological and Decision-Support Implications.

From an applied-mathematics standpoint, the framework contributes two decision-relevant capabilities. (i) It yields *design-consistent* predictive evaluation—weighted ROC/AUC, calibration, and bootstrap uncertainty—so that model performance reflects the target finite population, a prerequisite for downstream policy rules based on risk thresholds [38,39]. (ii) It provides *endogeneity-aware* causal effect estimates via orthogonal moments, ensuring that structural parameters used to justify escalation/escalation-avoidance policies are not artifacts of confounding. Together, these components support evidence-to-decision workflows that combine discrimination and calibration with explicit trade-offs, e.g., via decision-curve analysis and net benefit, which quantify whether model-guided actions improve outcomes relative to treat-all or treat-none strategies [42,43]. Practically, the pipeline informs when guideline adaptations (e.g., adding eGFR/ACR to general risk scores) produce measurable improvement in discrimination or reclassification [37], and when causal effects are too attenuated to warrant stand-alone threshold changes. The availability of mature survey software and workflows [1,40] facilitates reproducibility and technology transfer to public-health settings where resource allocation depends on design-consistent, decision-oriented metrics.

Limitations.

Although the framework is broadly applicable, several limitations must be recognized. First, causal interpretation of orthogonalized estimators still relies on strong identification assumptions—particularly approximate unconfoundedness or valid instruments—which cannot be empirically verified in observational surveys. Second, effective sample size can be severely limited for rare conditions like CKD, amplifying the variance of both predictive and causal estimators. Third, while the simulation study emulates key features of complex surveys, real-world surveys may present additional complications such as differential measurement error, multi-phase sampling, or uncon-

trolled response patterns. Fourth, the empirical application uses a unidirectional causal structure for conceptual clarity, though bidirectional cardio–renal interactions are clinically plausible. These limitations indicate that the empirical results should be viewed strictly as an illustrative case study validating the methodological pipeline rather than as substantive epidemiological findings.

Future Work.

Several extensions follow naturally from the present framework. One avenue concerns the development of survey-specific orthogonal learners that directly incorporate replicate weights or calibration constraints. Another concerns sharper finite-sample guarantees for penalized and debiased estimators under multi-stage sampling, where asymptotic approximations may be less reliable. A third direction is the integration of this framework into decision-support contexts, where prediction and causality must be balanced optimally under sampling constraints—e.g., combining design-consistent model evaluation with decision curves and cost–utility targets [42,43]. Finally, extending the theoretical analysis of design-consistent bootstrap methods for orthogonal estimators represents a promising path toward more rigorous foundations for causal inference in complex survey environments.

6. Conclusions

This study demonstrates the value of integrating survey-weighted inference, simulation analysis, and endogeneity-aware machine learning to separate predictive performance from causal interpretation in population health data. Using a nationally representative complex survey, we show that chronic kidney disease (CKD) is strongly associated with high predicted cardiovascular risk, yet its independent causal contribution is substantially attenuated once age and major cardiometabolic comorbidities are accounted for.

From a methodological standpoint, three lessons emerge. First, under complex sampling, predictive and causal targets need not coincide: penalized, survey-weighted models can retain strong discrimination and calibration for risk stratification, while orthogonalized estimators (e.g., DML) appropriately shrink structural effects in the presence of confounding and endogeneity. Second, survey-consistent performance assessment—weighted ROC/AUC, calibration, and design-preserving bootstrap—matters for population-level claims and prevents overstating model utility. Third, control-function and orthogonal methods provide complementary routes to causal identification; when assumptions hold, they recover causal parameters without sacrificing design awareness.

The simulation study confirms that naive predictive models may overstate causal effects under realistic confounding structures, whereas DML and control-function approaches can recover causal parameters when identification assumptions hold. Together, the empirical and simulation results underscore the importance of aligning modeling strategies with inferential objectives in complex surveys and of making explicit the trade-offs between prediction and causality.

Practically, the framework offers a scalable, design-consistent blueprint for evaluating extensions to cardiovascular risk tools (e.g., adding kidney measures) and for supporting decision-making with metrics that reflect the finite population and quantify uncertainty. Future work will (i) extend the analysis to additional survey waves and settings to assess transportability; (ii) develop survey-adapted orthogonal learners that directly incorporate replicate weights and calibration constraints; and (iii) further integrate decision-analytic tools (e.g., net benefit and cost–utility) with design-aware evaluation to optimize guideline adaptations and clinical decision support in CKD-specific cardiovascular risk management.

Supplementary Materials: The following supporting information can be downloaded at: [Mendeley ENS 2016–2017](#). Simulation and empirical study codes will be available soon at Zenodo (DOI to be provided upon publication).

Author Contributions: Conceptualization, Fernando Rojas and Axa Tapia; methodology, Axa Tapia and Fernando Rojas; software, Axa Tapia; validation, Fernando Rojas and Hilda Espinoza; formal analysis, Axa Tapia; investigation, Axa Tapia; resources, Hilda Espinoza; data curation, Axa Tapia; writing—original draft preparation, Axa

Tapia; writing—review and editing, Fernando Rojas and Hilda Espinoza; visualization, Axa Tapia; supervision, Fernando Rojas; project administration, Fernando Rojas; funding acquisition, Fernando Rojas. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Master's Program in Clinical Analysis, School of Chemistry and Pharmacy, Faculty of Pharmacy, Universidad de Valparaíso. The APC was funded by the same program.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to its nature as operational research focused exclusively on institutional logistics processes, without involving patient-level interventions or sensitive personal data.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data and code necessary to reproduce the results of this study are available in a public repository at Zenodo (<https://doi.org/10.5281/zenodo.19358307>). Simulated datasets and computational scripts are fully provided. The empirical analysis relies on data from the Chilean National Health Survey (ENS) 2016–2017, which are subject to data access policies and are therefore not redistributed. These data can be accessed through official channels of the Chilean Ministry of Health (MINSAL). The repository includes complete instructions to replicate the simulation and empirical analyses, [33].

Acknowledgments: The authors acknowledge the support of the academic staff of the Master's Program in Master's Program in Clinical Analysis for their guidance throughout the project.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lumley, T. *Complex Surveys: A Guide to Analysis Using R*; John Wiley & Sons: Hoboken, NJ, 2010.
2. Rao, J.N.K.; Wu, C.F.J. Resampling inference with complex survey data. *Journal of the American Statistical Association* **1988**, *83*, 231–241. <https://doi.org/10.1080/01621459.1988.10478591>.
3. Rust, K.; Rao, J. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **1996**, *5*, 283–310. <https://doi.org/10.1177/096228029600500305>.
4. Roy, J.; Shou, H.; Xie, D.; Hsu, J.Y.; Yang, W.; Anderson, A.H.; Landis, J.R.; Jepson, C.; He, J.; Liu, K.D.; et al. Statistical methods for cohort studies of CKD: prediction modeling. *Clinical Journal of the American Society of Nephrology* **2017**, *12*, 1010–1017. <https://doi.org/10.2215/CJN.06210616>.
5. Wang, J. Weighted estimation for multivariate shared frailty models for complex surveys. *Lifetime data analysis* **2019**, *25*, 469–479. <https://doi.org/10.1007/S10985-019-09469-X>.
6. Wooldridge, J.M. Control Function Methods in Applied Econometrics. *Journal of Human Resources* **2015**, *50*, 420–445. <https://doi.org/10.3368/jhr.50.2.420>.
7. Iparraguirre, J.; Lumley, T.; Barrio, I.; Arostegui, I. Variable selection with LASSO regression for complex survey data. *Stat* **2023**, *12*, e578. <https://doi.org/10.1002/sta4.578>.
8. McConville, K.S.; Breidt, F.J.; Lee, T.; Moisen, G.G. Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology* **2017**, *5*, 131–158. <https://doi.org/10.1093/jssam/smw041>.
9. Jasiak, J.; Tuvaandorj, P. Penalized Likelihood Inference with Survey Data, 2023, [arXiv:stat.ME/2304.07855]. <https://doi.org/10.48550/arXiv.2304.07855>.
10. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* **2018**, *21*, 1–68. <https://doi.org/10.1111/ectj.12097>.
11. Morzywolek, P.; Decruyenaere, J.; Vansteelandt, S. On Weighted Orthogonal Learners for Heterogeneous Treatment Effects. *arXiv preprint arXiv:2303.12687* **2024**.
12. Weiner, D.; et al. The Framingham Predictive Instrument in Chronic Kidney Disease. *Journal of the American College of Cardiology* **2007**, *50*, 217–224. <https://doi.org/10.1016/j.jacc.2007.03.037>.
13. Li, J.; et al. Performance of Cardiovascular Risk Scores in Patients with Chronic Kidney Disease: A Systematic Review. *American Journal of Kidney Diseases* **2022**, *79*, 689–703. <https://doi.org/10.1053/j.ajkd.2021.09.012>.
14. Taylor, J.M.; Ankerst, D.P.; Andridge, R.R. Validation of biomarker-based risk prediction models. *Clinical cancer research* **2008**, *14*, 5977–5983. <https://doi.org/10.1158/1078-0432.CCR-07-4534>.

15. Vasquez, M.M.; Hu, C.; Roe, D.J.; Chen, Z.; Halonen, M.; Guerra, S. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. *BMC medical research methodology* **2016**, *16*, 154. <https://doi.org/10.1186/s12874-016-0254-8>.
16. Matsushita, K.; et al. Association of estimated GFR and albuminuria with mortality. *The Lancet* **2010**, *375*, 2073–2081. [https://doi.org/10.1016/S0140-6736\(10\)60674-5](https://doi.org/10.1016/S0140-6736(10)60674-5).
17. Stenvinkel, P.; et al. Inflammation in End-Stage Renal Disease: The Hidden Enemy. *Nephrology Dialysis Transplantation* **2008**, *23*, 2133–2141. <https://doi.org/10.1093/ndt/gfn103>.
18. Matsushita, K.; et al. Inflammatory Markers and Cardiovascular Risk in Chronic Kidney Disease. *Kidney International* **2015**, *87*, 1104–1112. <https://doi.org/10.1038/ki.2014.403>.
19. Honda, H.; et al. Inflammation, Oxidative Stress, and Mortality in Hemodialysis Patients. *Kidney International* **2006**, *70*, 1486–1492. <https://doi.org/10.1038/sj.ki.5001752>.
20. Shoji, T.; et al. Oxidized LDL and Atherosclerosis in Chronic Kidney Disease. *Kidney International* **2001**, *59*, 1961–1967. <https://doi.org/10.1046/j.1523-1755.2001.00717.x>.
21. Shlipak, M.; et al. Cystatin C and prognosis for cardiovascular outcomes. *Annals of Internal Medicine* **2006**, *145*, 237–246. <https://doi.org/10.7326/0003-4819-145-4-200608150-00003>.
22. Peralta, C.; et al. Cystatin C Identifies Chronic Kidney Disease Patients at Higher Cardiovascular Risk. *Circulation* **2011**, *123*, 1382–1393. <https://doi.org/10.1161/CIRCULATIONAHA.110.983866>.
23. Herrington, W.; Staplin, N.; Judge, P.; et al. Evidence for Reverse Causality in the Association Between Blood Pressure and Cardiovascular Risk in Patients With Chronic Kidney Disease. *Hypertension* **2017**, *69*, 314–322. <https://doi.org/10.1161/HYPERTENSIONAHA.116.08386>.
24. Yang, W.; Wu, X.; Zhao, M.; et al., J.H. Kidney Function and Cardiovascular Disease: Evidence from Observational Studies and Mendelian Randomization Analyses. *Phenomics* **2024**, *4*, 250–253. <https://doi.org/10.1007/s43657-023-00145-7>.
25. Hu, C.; Li, Y.; Qian, Y.; et al., Z.W. Kidney Function and Cardiovascular Diseases: A Large-scale Observational and Mendelian Randomization Study. *Frontiers in Immunology* **2023**, *14*. <https://doi.org/10.3389/fimmu.2023.1190938>.
26. Terza, J.V.; Basu, K.; Rathouz, P.J. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* **2008**, *27*, 531–543. <https://doi.org/10.1016/j.jhealeco.2007.04.007>.
27. Terza, J.V. Two-Stage Residual Inclusion Estimation: A Practitioners Guide to Stata Implementation. *The Stata Journal* **2017**, *17*, 916–938.
28. Emmenegger, C.; Bühlmann, P. Regularizing Double Machine Learning in Partially Linear Endogenous Models. *Electronic Journal of Statistics* **2021**, *15*, 6461–6543. <https://doi.org/10.1214/21-EJS1931>.
29. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B* **1996**, *58*, 267–288.
30. Deo, R.; et al. Proteomic cardiovascular risk assessment in chronic kidney disease. *European Heart Journal* **2023**, *44*, 2095–2110. <https://doi.org/10.1093/eurheartj/ehad115>.
31. Levey, A.; Coresh, J. Chronic kidney disease. *The Lancet* **2012**, *379*, 165–180. [https://doi.org/10.1016/S0140-6736\(11\)60178-5](https://doi.org/10.1016/S0140-6736(11)60178-5).
32. Bansal, N.; et al. Cardiac Biomarkers and Risk of Incident Heart Failure in Chronic Kidney Disease. *Journal of the American Heart Association* **2019**, *8*. <https://doi.org/10.1161/JAHA.119.012336>.
33. Rojas, F. Reproducible Code and Data Repository: Design-Aware Predictive and Causal Modeling of Cardiovascular Risk in Chronic Kidney Disease Using Penalized and Double Machine Learning Approaches **2026**. <https://doi.org/10.5281/zenodo.19358307>.
34. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney International Supplement* **2024**, *105*.
35. Ndumele, C.E.; et al. A Synopsis of the Evidence for the Science and Clinical Management of Cardiovascular–Kidney–Metabolic (CKM) Syndrome. *Circulation* **2023**, *148*, 1636–1664. <https://doi.org/10.1161/CIR.0000000000001186>.
36. Hippisley-Cox, J.; Coupland, C.; Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease. *BMJ* **2017**, *357*, j2099. <https://doi.org/10.1136/bmj.j2099>.
37. Matsushita, K.; Kaptoge, S.; Hageman, S.H.J.; et al. Including measures of chronic kidney disease to improve cardiovascular risk prediction by SCORE2 and SCORE2-OP. In Proceedings of the European Heart Journal, 2022, Vol. 43, p. ehac544.2256. <https://doi.org/10.1093/eurheartj/ehac544.2256>.

38. Izrael, D.; Battaglia, A.A.; Hoaglin, D.C.; Battaglia, M.P. Use of the ROC Curve and the Bootstrap in Comparing Weighted Logistic Regression Models. In Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference (SUGI 27), 2002, pp. Paper 248–27.
39. Iparragirre, A.; Barrio, I.; Arostegui, I. svyROC: Estimation of the ROC Curve and the AUC for Complex Survey Data. CRAN package manual, version 1.0.0, 2025.
40. Feehan, D.M. *surveybootstrap: Bootstrap with Survey Data*, 2024. R package (dev) v0.1.0.90000.
41. Chernozhukov, V.; Hansen, C.; Kallus, N.; Spindler, M.; Syrgkanis, V. *Applied Causal Inference Powered by ML and AI*; 2024.
42. Vickers, A.J.; Elkin, E.B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **2006**, *26*, 565–574. <https://doi.org/10.1177/0272989X06295361>.
43. Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2 ed.; Springer: New York, 2019. <https://doi.org/10.1007/978-3-030-16399-0>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.