

A Two-Stage Detection Framework for Interaction Screening and Testing: A Systematic Simulation Evaluation of Method Suitability

Yue-peng Liu

liuyup2080@163.com

Saideng Biotech Co., Ltd.

Method Article

Keywords: Interaction detection, two-stage detection framework, method suitability, SHAP interaction values, multiple testing correction, bootstrap robustness, simulation study

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11505122/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background.

Interaction analysis faces two progressive demands, namely screening and testing. On one hand, p predictors give rise to $p(p - 1)/2$ potential interaction pairs; the number of candidates grows quadratically with dimensionality, and most of them are noise or spurious interactions, so efficient methods are needed to reduce the candidate space to a small set worthy of further examination. On the other hand, whether a candidate interaction reflects a real effect requires confirmation by methods with significance control. Existing interaction-analysis methods span different families, including parametric tests, semiparametric smoothing, model-based recursive partitioning, and machine-learning attribution, which differ markedly in capability positioning and failure modes. However, systematic evidence on which methods suit screening, which suit testing, and which suit neither is lacking. Using a "screen plus confirm" two-stage framework as the organizing principle, this study systematically evaluates the suitability of interaction-analysis methods and constructs an operational two-stage detection pipeline, clarifying its performance, scope of application, and robustness boundaries.

Methods.

Comparative experiments were conducted under three simulation scenarios (A, no interaction with collinear variables; B, linear multiplicative interaction; C, threshold-type nonlinear interaction), comparing single methods (product-term Wald test, GAM tensor-product interaction, Friedman H statistic, model-based recursive partitioning MOB, SHAP interaction top-1) with the two-stage method in end-to-end performance. The methods were then refined through successive rounds, including top- k screening-width evaluation, screener-family comparison (SHAP interaction screening vs. SIS-type marginal screening), Benjamini–Hochberg (BH) multiple-testing correction plus bootstrap robust confirmation, a robustness grid (collinearity ρ , sample size n , event rate, and outcome type), comparative experiments on collinearity preprocessing and screening-stage noise calibration, comparison of Wald vs. likelihood-ratio tests and of different bootstrap reproducibility thresholds at the confirmation stage, and a sensitivity–specificity validation of the MOB mechanism-interpretation channel. Finally, case analyses were conducted on two real datasets (the Dryad CRRT clinical cohort and the Dryad GMDR public data) to examine the behavior of the framework in real-world settings.

Results.

In the single-method comparison, SHAP interaction showed a false-positive rate as high as 64% in scenario A, while the product-term test achieved only 85% detection in scenario C, highlighting the necessity of the "screen–confirm" division of labor. The two-stage method (top-1 screening) reduced the end-to-end false-positive rate in scenario A to about 5% while maintaining 83% detection in scenario C. The top- k evaluation showed $k = 5$ to be the best compromise (99% coverage in scenario C). The screener-family comparison showed that the SIS-type statistical-test screener had slightly lower coverage in the threshold scenario (90%) and, because it is informationally homogeneous with the confirmation stage, drove the candidate-set false-confirmation rate in the no-interaction scenario up to 96% (vs. 58% for SHAP). BH plus bootstrap reproducibility ≥ 0.8 reduced the false-confirmation rate in scenario A from 58% to 15% and brought the end-to-end false-positive rate to zero; raising the bootstrap threshold to 0.9 further reduced the false-confirmation rates in scenarios A/B/C to 7%, 39%, and 8%, at the cost of lowering the robust detection rate in scenario C to 47%. The robustness grid showed that the framework is robust

to collinearity (52%–83% detection in scenario C for $p \in [0, 0.7]$) and has a clear sample-size threshold (detection in scenario C collapsed to 5% at $n = 250$ and became reliable only at $n \geq 500$), with detection unexpectedly higher under Cox survival outcomes (100%). Collinearity residualization was not recommended because it removed the true interaction signal (scenario C detection 58%→37%), and permutation noise calibration gave limited benefit (scenario B false confirmation 49%→40% but at 30-fold computational cost). Systematic validation of MOB as an independent evidence channel showed a target-pair capture rate of 58% for threshold-type interactions (62% end-to-end in parallel dual-channel mode), but a false-alarm rate as high as 26% on spurious pairs in the no-interaction scenario that partitioned-variable residualization could not suppress; MOB was therefore not used as independent confirmation evidence and was retained only as a mechanism-refinement tool for confirmed interactions. In the real-data case analyses, the framework screened $P_{24h} \times CCI$ out of 595 variable pairs in the CRRT cohort ($n = 1144$, 90-day mortality) and assigned it to Level 3 "preliminary confirmation" (bootstrap reproducibility 0.84, below the 0.9 threshold), and rejected the literature-reported but statistically unconfirmable ITGA2rs4865756×HABP2rs7923349 interaction in the GMDR data ($n = 2377$, carotid stenosis), validating the screen–confirm division of labor in real-world settings.

Conclusions.

The two-stage framework balances the trade-off between a high false-positive rate in screening and insufficient sensitivity in testing. The recommended final configuration is SHAP interaction top- k screening ($k = 5$) → product-term Wald test → BH correction → bootstrap reproducibility ≥ 0.9 robust confirmation, supplemented by a four-level evidence-grading framework (candidate indication, preliminary evidence, preliminary confirmation, robust confirmation). The mechanism-interpretation channel (MOB) was shown to have high sensitivity but uncontrolled false positives and serves only as a mechanism-refinement tool for confirmed interactions rather than independent confirmation evidence. The framework is applicable to binary or survival outcomes with sample size $n \geq 500$, event rate $\geq 30\%$, and collinearity strength $\rho \leq 0.7$; false-confirmation rates remain elevated when strong linear interactions coincide with strong collinearity, an inherent methodological challenge that should be addressed by combining multiple evidence sources. The real-data case analyses further corroborate the framework's dual capability of screening out candidates and rejecting spurious interactions.

1. Introduction

Interaction analysis holds important mechanistic value in clinical and epidemiological research, for example in identifying gene–environment interactions, joint effects of biomarkers, and treatment-effect heterogeneity; interaction effects often carry more direct information about the underlying pathways than main effects. However, interaction analysis faces two progressive demands. The first is screening. With p predictors there are $p(p - 1)/2$ potential interaction pairs; the number of candidates grows quadratically with dimensionality, and the vast majority are noise or spurious interactions, so the researcher first needs a method that can rapidly shrink the candidate space, at acceptable computational cost, to a small set of interaction pairs "worthy of further examination." The second is testing. Whether a shrunk candidate interaction truly exists and whether its effect is robust requires evidence with significance control and uncertainty quantification. The two demands impose requirements on methods along different dimensions—screening emphasizes coverage and computational efficiency, while testing emphasizes the stringency of significance control—and neither can be fully satisfied by a single method, constituting the fundamental tension in the choice of interaction-analysis methods.

Around these two demands, existing methods can be broadly divided by design purpose into "screening-oriented" and "testing-oriented" classes. Screening-oriented methods originate from the dimension-reduction ideas of high-dimensional variable selection. Sure independence screening (SIS) uses marginal correlations to rapidly reduce ultrahigh-dimensional features (Fan & Lv, 2008; Fan et al., 2009); its interaction-scenario extensions include TS-GSIS (Fang et al., 2017), EPISIS (Hibar et al., 2015), and the generic nonparametric BCor-SIS (Pan et al., 2019). Multifactor dimensionality reduction methods, such as GMDR, search for the best interaction model combinations by cross-validation consistency. On the machine-learning side, tree-ensemble models can fit complex interaction structures without a prespecified functional form; their attribution tools, such as SHAP interaction values (Lundberg et al., 2018; Lundberg et al., 2020) and the Friedman H statistic (Friedman & Popescu, 2008), are widely used for batch screening of candidate interactions, as are the regional effect plots REPID (Herbinger et al., 2022) and the permutation-test-based GGInt-XGBoost (Guo et al., 2021). Testing-oriented methods center on significance testing. Parametric models provide p values and effect sizes through Wald or likelihood-ratio tests on product-term coefficients (but assume a linear multiplicative form and have low power for nonlinear interactions such as threshold-type effects); semiparametric methods characterize interactions with smoothing terms and test them by likelihood ratios, such as GAM tensor-product terms; model-based recursive partitioning (MOB) identifies structural changes via parameter-stability tests (Zeileis et al., 2008); and permutation tests provide significance evidence directly from empirical distributions. Overall, screening methods have strong coverage but lack control over significance and tend to produce systematic spurious interactions when variables are collinear in the absence of true interactions; testing methods have strict significance control but are sensitive to the functional form of the interaction and have limited power under nonlinear forms. The two classes belong to different paradigms with complementary strengths and weaknesses, yet existing research has mostly developed within a single paradigm. Systematic evidence evaluating method suitability by "screen-test" role within a common simulation framework is lacking, as is quantitative characterization of the end-to-end performance of two-stage combinations (engineering parameters such as screening width, multiple-testing correction, robustness thresholds, and preprocessing choices).

Adopting "AI-assisted screening + simulation with known truth + public-data validation" as the methodological framework, this study systematically evaluates the suitability of methods within the interaction-analysis family around the two-stage demand of "screening plus confirmation"—which methods are suited to candidate convergence at the screening stage, which to significance testing at the confirmation stage, and which to neither—and accordingly constructs an operational two-stage interaction-detection pipeline. The main contributions are threefold. (1) The failure modes of five representative methods and the performance difference of the two-stage pipeline are quantified under three simulation scenarios, providing evidence for the necessity of the "screen-confirm" division of labor. (2) Key parameters of the two-stage pipeline are refined item by item, including screening width, BH multiple-testing correction and bootstrap robust confirmation, a four-dimensional robustness grid, and comparative experiments on collinearity preprocessing and noise calibration. (3) An operational evidence-grading framework, method-suitability conclusions, and a scope-of-application statement are established, and the sensitivity–specificity boundary of the mechanism-interpretation channel (MOB) is empirically validated, providing methodological grounds for the use of such methods on real data.

2. Methods

2.1 Problem setting and notation

Let the data be (X, Y) , where $X \in \mathbb{R}^{n \times p}$ denotes p predictors and $Y \in \{0,1\}$ is a binary outcome (or (T, δ) denotes a survival outcome). The interaction effect of a variable pair is defined as the dependence of the outcome on the joint variation of the variables beyond what is explained by main effects. The detection target of this study is the set of true interaction pairs among all $p(p-1)/2$ variable pairs. Let $\varphi_{ij}(x_\ell)$ denote the SHAP interaction value for variables i and j on the ℓ -th individual (symmetrically averaged), and define the global interaction strength as

$$S_{ij} = n^{-1} \sum_{\ell} (|\varphi_{ij}(x_\ell)| + |\varphi_{ji}(x_\ell)|)$$

2.2 Two-stage detection framework

The framework consists of a screening stage and a confirmation stage. The screening stage trains an XGBoost model (binary logistic or survival:cox objective), computes the SHAP interaction strength S_{ij} for all variable pairs, and takes the top- k pairs in descending order of strength to form the candidate set C . The confirmation stage fits, for each candidate pair (i, j) , a logistic (or Cox) model containing all main effects and the product term $X_i \cdot X_j$, using the Wald-test p value of the product-term coefficient as preliminary significance evidence; applies Benjamini–Hochberg (BH) false-discovery-rate (FDR) correction to the p values in the candidate set; and, for candidate pairs passing the correction, performs B bootstrap resampling repetitions of the test and computes the reproducibility ρ_{ij} (the proportion of resamples in which the test is significant). Algorithm 1 gives the complete procedure.

Algorithm 1

Two-stage interaction detection procedure (final configuration).

Input

data (X, Y) ; screening width k ; number of bootstrap resamples B ; reproducibility threshold τ

Output

candidate indication set C ; preliminary confirmation set P ; robust confirmation set R

1. 1. Train XGBoost model $\mathcal{F}$, compute SHAP interaction values $\varphi_{ij}(x_\ell)$, $\ell=1, \dots, n$.
2. 2. Compute global strengths S_{ij} (for all variable pairs).
3. 3. $C \leftarrow$ top- k variable pairs by descending S_{ij} .
4. 4. for $(i, j) \in C$:
5. 5. Fit a model with all main effects and the product term $X_i \cdot X_j$.
6. 6. $p_{ij} \leftarrow$ Wald-test p value of the product term.
7. 7. $P \leftarrow \{ (i, j) \in C : p_{ij} < 0.05 \text{ after BH correction} \}$.
8. 8. for $(i, j) \in P$:
9. 9. $\rho_{ij} \leftarrow$ proportion of B bootstrap resamples in which the test is significant.
10. 10. $R \leftarrow \{ (i, j) \in P : \rho_{ij} \geq \tau \}$.
11. 11. return C, P, R .

2.3 Simulation framework

The simulation data are generated as follows. Predictor X_1 follows a standard normal distribution, X_2 is generated as $X_2 = \rho X_1 + \sqrt{(1 - \rho^2)}\epsilon$ to mimic collinearity with X_1 (default $\rho = 0.5$), and the remaining eight variables correlate

with X_1 at 0.3. The three scenarios correspond to three kinds of "known truth." Scenario A (no interaction, collinear) contains only main effects and is used to evaluate false positives. Scenario B (linear multiplicative interaction) adds $0.8 \cdot X_1 \cdot X_2$ to the main effects, with the interaction in linear form. Scenario C (threshold-type nonlinear interaction) adds $0.8 \cdot X_2 \cdot I(X_1 > 0.5)$ to the main effects, with the interaction bounded by a threshold on X_1 —a typical nonlinear form that a linear product term cannot express directly. Outcome probabilities are given by a logistic link, with the intercept iteratively calibrated so that the event rate is approximately 50% (specified separately in the robustness grid). XGBoost parameters are fixed at $\eta = 0.1$, $\text{max_depth} = 4$, and $\text{nrounds} = 100$. The default settings are $n = 1000$, $p = 10$, $\rho = 0.5$, and an event rate of 50%, with 100 independent replicates per cell (60 per cell in the robustness grid) and parallel random seeds for reproducibility. All simulations were implemented in R 4.5.3 using the `xgboost`, `mgcv`, `partykit`, `survival`, and `parallel` packages. Table 1 summarizes the scenario settings.

Table 1
Simulation scenario settings ($n = 1000$, $p = 10$, $\rho = 0.5$, event rate $\approx 50\%$, 100 replicates).

Scenario	Outcome generation	True interaction	Purpose
A no-interaction collinear	$\text{logit}(\beta_0 + 0.25\sum X)$	None	Evaluate false positives / false confirmation
B linear multiplicative interaction	$\text{logit}(\beta_0 + 0.25\sum X + 0.8 \cdot X_1 \cdot X_2)$	$X_1 \times X_2$ (linear)	Evaluate linear-interaction detection
C threshold-type nonlinear interaction	$\text{logit}(\beta_0 + 0.25\sum X + 0.8 \cdot X_2 \cdot I(X_1 > 0.5))$	$X_1 \times X_2$ (threshold)	Evaluate nonlinear-interaction detection

2.4 Comparator methods

The single-method comparison includes five representative methods. The product-term Wald test (`glm`) fits a logistic regression containing all main effects and the target product term and tests the product-term coefficient. The GAM tensor-product interaction term (`gam ti`) characterizes the interaction with a tensor-product smooth term $ti(X_i, X_j)$ and tests its significance with a likelihood-ratio test. SHAP interaction top-1 directly reports the pair with the highest global SHAP interaction strength as the detection. The Friedman H statistic (H top-1) computes the Friedman–Popescu interaction strength H_{ij} and takes the highest pair. Model-based recursive partitioning (MOB) recursively partitions the data by parameter-stability tests and records whether a structure with a candidate variable as the split point appears. Both MOB and GAM include all main effects in the model as a modeling discipline, avoiding the false-positive inflation caused by omitting main effects. For the end-to-end evaluation of the two-stage method, detection is defined as successful confirmation within the candidate set, and false confirmation is defined as a spurious pair within the candidate set passing confirmation in the no-interaction scenario.

The choice of comparator methods follows the principle of paradigm representativeness. The five methods represent five major statistical paradigms—parametric testing (product-term Wald), semiparametric smoothing (GAM `ti`), machine-learning attribution (SHAP), nonparametric interaction statistics (Friedman H), and model-based recursive partitioning (MOB)—which suffice to characterize the suitability division of "heavy coverage at the screening stage and heavy control at the confirmation stage." In addition, an SIS-type marginal screener (v12, ranking by pairwise "all main effects plus product term" Wald-test p values) was set up as a within-family comparator for the screening stage, to examine the relative suitability of "statistical-test ranking vs. model-attribution ranking." Some method families introduced in the introduction were not included in the replicate-by-replicate simulation comparison, for the following reasons. The SIS lineage (TS-GSIS, EPISIS, BCor-SIS) achieves

ranking-based shrinkage through marginal correlations or marginal tests, and its interaction extensions mainly target genome-wide ultrahigh-dimensional settings, where they are not representative at the simulation dimensionality $p = 10$; this study implements the ranking-shrinkage mechanism of the family via Wald-test ranking and compares it with SHAP. GMDR, which centers on cross-validation consistency and sign tests with computational cost growing quadratically in the number of variables, is too expensive for replicate-by-replicate simulation; its empirical case is instead used as the real-data re-examination object (see § 3.9.2 and § 4.2). The permutation-test idea of GGInt-XGBoost is covered by the v8 permutation noise-calibration comparison (see § 3.6).

2.5 Evaluation metrics

The following metrics are defined. The screening coverage rate is the proportion of true interaction pairs entering the candidate set C . The end-to-end detection rate is the proportion of true interaction pairs still reported after BH correction and the bootstrap reproducibility threshold. The false-confirmation rate is the proportion of spurious pairs passing each confirmation stage in the no-interaction scenario. The all-pairs testing baseline fits, without screening, a model containing all interaction terms for all 45 variable pairs simultaneously and reports the significant ones, quantifying the multiple-testing burden. All metrics are averaged over 100 independent replicates (60 for the grid).

2.6 Refinement process of the framework

To evaluate the influence of the framework's engineering parameters, nine comparative experiments were carried out in sequence. (1) Single-method comparison (v2, five methods; the early three-method exploratory comparison was not included in the formal results). (2) End-to-end evaluation of the two-stage method (v3, screening width $k = 1$). (3) Top- k screening-width evaluation (v4, $k \in \{1, 3, 5, 10, 15\}$). (4) BH correction and bootstrap robust confirmation at the confirmation stage (v5, $B = 50$, reproducibility threshold 0.8). (5) Robustness grid (v6, $p \in \{0, 0.3, 0.5, 0.7\} \times n \in \{250, 500, 1000\} \times \text{event rate} \in \{10\%, 30\%, 50\% \} \times \text{outcome} \in \{\text{binary}, \text{Cox}\}$, 60 replicates per cell). (6) Collinearity-preprocessing comparison (v7, no preprocessing vs. residualization of X_2) and screening-stage permutation noise-calibration comparison (v8, constructing a null strength distribution by permuting Y 30 times per replicate, filtering by $p_{\text{perm}} < 0.05$). (7) Confirmation-stage upgrade comparison (v9, Wald vs. likelihood-ratio test $\times$ reproducibility threshold 0.8 vs. 0.9). (8) Validation of the MOB mechanism-interpretation channel (v10, binary focused MOB + candidate-set BH + bootstrap significance-reappearance gate; v11, partitioned-variable residualization control comparison). (9) Screener-family comparison (v12, SIS-type pairwise-test screening vs. SHAP interaction screening).

2.7 Evidence-grading framework

Based on the simulation results, interaction evidence is organized into four levels. Level 1 "candidate indication" refers to entering the SHAP top- k candidate set. Level 2 "preliminary evidence" refers to a product-term Wald-test $p < 0.05$ without passing BH correction. Level 3 "preliminary confirmation" refers to passing BH correction with a bootstrap reproducibility below 0.9. Level 4 "robust confirmation" refers to a bootstrap reproducibility ≥ 0.9 after BH correction. Mechanism interpretation is an independent dimension applied only to interaction pairs that have passed BH correction (Level 3 and above), where model-based recursive partitioning (MOB) reports split points and stratified effect sizes to provide the form and cut point of the interaction; this dimension does not participate in the end-to-end detection decision. The framework presents the three layers of evidence—screening, correction, and robustness—separately, avoiding overinterpretation of a single statistic, and performs mechanism interpretation after statistical confirmation, avoiding the substitution of structural evidence for significance evidence.

2.8 Real-data validation design

To examine the behavior of the framework on real data, two datasets were selected for case analysis. (1) CRRT cohort. A public dataset from the Dryad Digital Repository (Jung et al., 2018; Jung et al., 2019), whose source study is a retrospective cohort on phosphate as a biomarker predicting adverse outcomes in AKI patients undergoing CRRT (PLoS ONE). It contains $n = 1144$ patients with acute kidney injury receiving continuous renal replacement therapy (CRRT), with the outcome of 90-day all-cause mortality (Cox, 821 events, event rate 71.8%) and 31 predictors (18 continuous + 10 binary + 3 multicategory; missing values were imputed with MICE, 5 imputations, first set used). (2) GMDR data. A public dataset from the Dryad Digital Repository (Liu, 2022; Liu et al., 2023), with $n = 2377$ individuals at high risk of stroke, the outcome of carotid stenosis (binary, 295 events, event rate 12.4%), and predictors consisting of sex and six inflammation/endothelial-function-related SNPs (two loci each for IL1A, ITGA2, and HABP2). The original paper reported ITGA2rs4865756×HABP2rs7923349 as a significant gene–gene interaction by generalized multifactor dimensionality reduction (GMDR) (permutation test $P = 0.010$, cross-validation consistency 8/10, sign test 10/10, adjusted OR = 1.42 for the high-risk interacting genotype), and this study uses it as a "known truth" comparator. The analysis protocol is identical to the simulation. XGBoost was trained ($\eta = 0.1$, $\text{max_depth} = 4$, $\text{nrounds} = 100$), SHAP interaction strengths were computed, the top-5 candidates were taken, a logistic (or Cox) model with all main effects and the product term was fitted for each candidate with a Wald test and BH correction, and $B = 100$ bootstrap reproducibility tests (threshold 0.9) were performed for those passing. For the GMDR data, a factor-by-factor (3×3 genotype combinations, 4 degrees of freedom) likelihood-ratio test was additionally used as a sensitivity analysis close to the original GMDR metric; for the CRRT data, restricted cubic splines (rcs, 4 knots) for main effects instead of linear main effects were additionally used as a sensitivity analysis.

3. Results

3.1 Single-method comparison

The detection rates of the five methods under the three scenarios are shown in Table 2. In scenario A (no interaction, collinear), the SHAP interaction top-1 false-positive rate was as high as 64% and that of Friedman H was 30%, while the three statistical-test methods (glm 6%, GAM 7%, MOB 9%) stayed near the nominal level. In scenario B (linear multiplicative), all methods detected close to 100%. In scenario C (threshold-type nonlinear), SHAP interaction detected 98%, while the linear product-term test detected only 85%, GAM 81%, Friedman H 65%, and MOB 57%. This comparison shows that the failure modes of the two paradigms are complementary: SHAP has high false positives under collinearity without interaction, and statistical tests have insufficient sensitivity to nonlinear interactions; the two require division of labor rather than substitution.

Table 2
Single-method detection rates (v2, 100 replicates; values for scenario A are false-positive rates).

Scenario	glm product term	GAM ti	SHAP top-1	Friedman H top-1	MOB
A no-interaction collinear	0.06	0.07	0.64	0.30	0.09
B linear multiplicative	1.00	1.00	1.00	0.99	1.00
C threshold nonlinear	0.85	0.81	0.98	0.65	0.57

3.2 End-to-end evaluation of the two-stage method

With SHAP top-1 screening followed by glm/GAM/MOB confirmation (v3), the end-to-end results are shown in Table 3. In scenario A, the false-positive rate dropped from 64% with SHAP alone to 5% (glm)/5% (GAM)/7% (MOB)

with the two-stage method, with false-confirmation rates of only 4%–7%. In scenario C, the end-to-end detection rate was maintained at 83% (glm)/81% (GAM), while the MOB confirmer had lower sensitivity (55%). These results show that the two-stage design reduces the false-positive rate by an order of magnitude without losing nonlinear-interaction detection, and that the choice of confirmer has limited influence within the family of statistical tests.

Table 3
End-to-end evaluation of the two-stage method (v3, SHAP top-1 screening; scenario A values are end-to-end false positives).

Scenario	SHAP alone	All-pairs baseline	+glm confirm	+GAM confirm	+MOB confirm
A no-interaction collinear	0.64	0.06	0.05	0.05	0.07
B linear multiplicative	1.00	1.00	1.00	1.00	1.00
C threshold nonlinear	0.98	0.85	0.83	0.81	0.55

3.3 Choice of screening width k

The top-*k* evaluation (v4) is shown in Table 4. In scenario C, the coverage of the true interaction pair in the candidate set increased with *k*, from 91% at *k* = 1 to 98% at *k* = 3, 99% at *k* = 5, and 100% at *k* = 10; in scenario A, the coverage of spurious pairs rose in parallel (37% at *k* = 5), indicating that a wider screen comes at the cost of more false candidates. In scenario C, the end-to-end detection rate reached 82% at *k* = 5 (Table 4), and the all-pairs testing baseline (without screening) detected only 68% while reporting on average about 2–3 false-positive pairs, quantifying the multiple-testing burden. Considering coverage, end-to-end detection, and computational cost, *k* = 5 is the best compromise. Although *k* = 1 had the fewest false candidates (13% in scenario A), its coverage of only 91% entails a higher risk of missing true interactions.

To answer which of tree-model attribution screening and statistical-test screening is better suited to the screening stage, an SIS-type marginal screener (pairwise "all main effects plus product term" Wald-test *p* values ranked ascending and top-*k* taken, v12) was compared within the same framework (lower half of Table 4). The results are as follows. In the linear-interaction scenario (B), the two were fully equivalent (coverage 1.00, end-to-end 1.00). In the threshold-type nonlinear scenario (C), SIS had slightly lower coverage (90% at *k* = 5 vs. 99% for SHAP) but identical end-to-end detection (82% vs. 82%), because the target pairs entering the candidate set were also mostly significant under the linear test. The core difference lay in the no-interaction scenario (A). SHAP ranked the specific spurious pair induced by the collinear structure ($X_1 \times X_2$) into top-5 in 37% of replicates, versus only 16% for SIS, close to the random level of $5/45 = 11\%$ (i.e., not biased toward the collinear structure); however, SIS had a candidate-set false-confirmation rate as high as 96% (vs. 58% for SHAP). The mechanism of this difference lies in the informational homogeneity between the screener and the confirmer. SIS ranks candidates by interaction-test *p* values, and the confirmation stage admits the same test, so the spurious pair with the smallest *p* value is inevitably confirmed, sending nearly the full multiple-testing burden into the confirmation stage. SHAP, by contrast, ranks by model attribution, which is orthogonal to the statistical test and provides the confirmation stage with an independent information dimension. This comparison methodologically supports the principle that "the screening stage should select a ranking signal informationally heterogeneous from the confirmation stage," and forms part of the suitability conclusions.

Table 4

Screeners-family top- k coverage and end-to-end performance (v4 SHAP / v12 SIS, 10 variables, 45 candidate pairs; end-to-end and candidate-set false confirmation are $k = 5$ configurations).

Screeners	Scenario	Rank med.	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 15$	End-to-end ($k = 5$)	Cand.-set false conf. ($k = 5$)	All-pairs baseline	All-pairs FP median
SHAP	A no-interaction	7	0.13	0.27	0.37	0.60	0.68	0.02	0.58	0.05	2
SHAP	B linear	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	4
SHAP	C threshold	1	0.91	0.98	0.99	1.00	1.00	0.82	0.52	0.68	3
SIS	A no-interaction	19.5	0.02	0.08	0.16	0.29	0.38	0.05	0.96	0.05	2
SIS	B linear	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	4
SIS	C threshold	1	0.63	0.83	0.90	0.96	0.97	0.82	0.90	0.68	3

3.4 FDR control and robustness at the confirmation stage

The confirmation-stage upgrade comparison (v5, $k = 5$, $B = 50$) is shown in Table 5. In scenario A, the in-candidate-set false-confirmation rate fell from 58% without correction to 25% after BH and further to 15% after adding bootstrap (reproducibility ≥ 0.8), and the end-to-end false-positive rate (already only 2% without correction) reached zero after BH. The corresponding reductions were 90% $\rightarrow$ 80% $\rightarrow$ 49% in scenario B and 52% $\rightarrow$ 33% $\rightarrow$ 13% in scenario C. The median bootstrap reproducibility of true interaction pairs was 100% in scenario B and 94% in scenario C, indicating that robust confirmation did not sacrifice true signals. The elevated false-confirmation rate in scenario B (strong linear interaction) arises because, under the collinear structure, SHAP ranks weak spurious interactions into the top-5 and their p values remain stably significant under resampling; this issue is further analyzed by the screening and preprocessing comparisons.

Table 5

Confirmation-stage BH and bootstrap upgrade (v5, $k = 5$, $B = 50$, reproducibility ≥ 0.8).

Scenario	Coverage	End-to-end raw	End-to-end + BH	End-to-end + BS	False conf. raw	False conf. +BH	False conf. +BS	Target-pair BS reprod. median
A no-interaction	0.37	0.02	0.00	0.00	0.58	0.25	0.15	—
B linear	1.00	1.00	1.00	1.00	0.90	0.80	0.49	1.00
C threshold	0.99	0.82	0.68	0.58	0.52	0.33	0.13	0.94

3.5 Robustness grid

The four-dimensional robustness grid (v6) results are shown in Table 6. Along the collinearity dimension, the robust detection rate in scenario C declined slowly from 83% to 52% as ρ increased from 0 to 0.7, with the false-confirmation rate always controlled (10%–20% after bootstrap) and end-to-end false positives in the no-interaction scenario at 0–3.3% across all grid points, showing that the method's conclusions do not depend on a particular

collinearity level. The sample-size dimension shows a clear threshold: the robust detection rate in scenario C collapsed to 5% at $n = 250$, was 28% at $n = 500$, and was 63% at $n = 1000$, while SHAP coverage rose in parallel from 60% to 98% (Table 6), indicating that both screening and confirmation fail simultaneously at small sample sizes. Along the event-rate dimension, the detection rate was only 38% at the sparse event rate of 10% and stabilized at 75% and 63% at 30% and 50%. Along the outcome-type dimension, Cox survival outcomes were unexpectedly better than binary outcomes (robust detection in scenario C 100% vs. 63%, coverage 100%), showing that the method extends naturally to survival data.

Table 6

Key cells of the robustness grid (v6, 60 replicates per cell; C detection = scenario C robust detection, A false confirmation = scenario A end-to-end false confirmation, C coverage = proportion of scenario C target pairs entering top-5).

Dimension	Levels	C detection	A false confirmation	C coverage
ρ ($n = 1000$, 50%)	0 / 0.3 / 0.5 / 0.7	0.83 / 0.82 / 0.63 / 0.52	0.13 / 0.20 / 0.13 / 0.10	0.98 / 1.00 / 0.98 / 1.00
n ($\rho = 0.5$, 50%)	250 / 500 / 1000	0.05 / 0.28 / 0.63	0.13 / 0.12 / 0.13	0.60 / 0.95 / 0.98
Event rate ($n = 1000$)	10% / 30% / 50%	0.38 / 0.75 / 0.63	0.15 / 0.10 / 0.13	1.00 / 1.00 / 0.98
Outcome ($n = 1000$)	Binary / Cox	0.63 / 1.00	0.13 / 0.15	0.98 / 1.00

3.6 Collinearity preprocessing and noise calibration

The results of the collinearity-residualization comparison (v7) and the permutation noise-calibration comparison (v8) are shown in Table 7. After residualizing X_2 on all other variables, the robust detection rate in scenario C fell from 58% to 37%, the false-confirmation rate in scenario B decreased only slightly from 49% to 44%, and the coverage in scenario A fell from 37% to 22% without improvement in the false-confirmation rate. The reason is that part of the threshold-type interaction signal is carried by the collinear component of X_1-X_2 , which orthogonalization removes, while false confirmation mainly stems from a systematic bias of SHAP across the entire correlation structure that single-variable orthogonalization cannot eliminate. The screening-stage calibration that constructed a null strength distribution by permuting Y (30 permutations per replicate) reduced the false-confirmation rate in scenario B from 49% to 40% and in scenario A from 15% to 12%, while detection in scenario C remained unchanged at 58%; however, the gain came at 30-fold computational cost, and because the permutation null tests "existence of association" rather than "interaction specificity," it cannot filter out collinear spurious interactions transmitted through main effects. Together, the two comparisons show that the effective treatment of collinearity lies not in data preprocessing but in strict standards at the confirmation stage.

Table 7
Preprocessing and noise-calibration comparison (v7/v8, 100 replicates; the no-preprocessing row is the v5-configuration baseline).

Scheme	C robust detection	B false confirmation	A coverage	A false confirmation
No preprocessing (v5 configuration)	0.58	0.49	0.37	0.15
X_2 residualization (v7)	0.37	0.44	0.22	0.18
Permutation noise calibration (v8)	0.58	0.40	0.30	0.12

3.7 Confirmation-stage upgrade and final configuration

The four-variant confirmation-stage comparison (v9) is shown in Table 8. Raising the bootstrap reproducibility threshold from 0.8 to 0.9 reduced the false-confirmation rates in scenarios A/B/C from 15%/49%/13% to 7%/39%/8%. Replacing the Wald test with the likelihood-ratio (LR) test made almost no difference when used alone (lr08 and wald08 were indistinguishable within noise across all metrics), and yielded only a marginal 3-percentage-point improvement in scenario B (39%→36%) when combined with the 0.9 threshold, at the cost of a further 4-percentage-point drop in the robust detection rate in scenario C (47%→43%). The recommended final configuration is Wald + reproducibility threshold 0.9 (wald09). Under this configuration, scenario A false confirmation is 7%, scenario B 39%, scenario C 8%, robust detection in scenario C is 47%, and scenario B remains at 100%. The sensitivity loss is the inevitable price of a stricter evidence standard and is suggested to be compensated by the mechanism-interpretation dimension (MOB segmentation rules and similar; feasibility validated in § 3.8).

Table 8
Four confirmation-stage variants (v9, 100 replicates; C detection = scenario C robust detection).

Variant	A false confirmation	B false confirmation	C false confirmation	C detection
wald08 (Wald, BS $\geq$ 0.8)	0.15	0.49	0.13	0.58
lr08 (LR, BS $\geq$ 0.8)	0.16	0.49	0.17	0.61
wald09 (Wald, BS $\geq$ 0.9)	0.07	0.39	0.08	0.47
lr09 (LR, BS $\geq$ 0.9)	0.07	0.36	0.08	0.43

3.8 Simulation evaluation of the MOB mechanism-interpretation channel

To examine the feasibility of "supplementing mechanism interpretation with MOB," the performance of MOB as an independent evidence channel was systematically validated (v10, 100 replicates; Table 9). The implementation used a binary focused version of model-based recursive partitioning (the mob framework in partykit with binomial logistic regression and the parm argument restricting the test to changes in the target variable's coefficient with the partitioning variable, avoiding the dilution of the full-parameter joint test by stable scores of irrelevant variables at $p = 10$). For each candidate pair, bidirectional partitioning tests were performed (X_j partitioning focusing on the X_i coefficient $\cup$ X_i partitioning focusing on the X_j coefficient), significance was corrected by candidate-set BH, and stability was judged by a bootstrap significance-reappearance rate ≥ 0.6 . The results are as follows. In scenario C (threshold interaction), the target-pair capture rate was 58%, and the parallel dual-channel end-to-end detection rate

rose to 62% (a 15-percentage-point gain relative to the 47% of the confirmation stage alone). However, in scenario A (no interaction, collinear), the false-alarm rate on spurious pairs was as high as 26%, and the nominal calibration of the single-test p values was loose (about 12% with $p < 0.05$ vs. the nominal 5%). Two further attempts at collinearity control both failed (Table 9). Partitioned-variable residualization (v11) did not reduce false positives (23% vs. 26%) and substantially lost sensitivity (target-pair capture 34% vs. 58%). A split-point location stability gate (proportion of bootstrap split points falling within the ± 0.2 bandwidth ≥ 0.6) reduced false positives to 10% but at the cost of halving sensitivity (24%). These results show that the MOB test has high sensitivity to threshold-type interactions, but its false positives come mainly from collinear spurious pairs and are structurally stable—they reappear repeatedly under resampling and cannot be eliminated simultaneously by orthogonalization or a location gate. Accordingly, MOB is not incorporated into the end-to-end decision as independent confirmation evidence; it retains only its mechanism-interpretation role, namely reporting split points and segment effects for confirmed interactions to provide a structured reading of the interaction form.

<div>Table 9</div> <div>Validation of the MOB mechanism-interpretation channel (v10/v11, 100 replicates each; C capture = scenario C target-pair capture rate, A false alarm = scenario A non-target candidate false-alarm rate).</div>			
Scheme	C capture	A false alarm	Description
v10 final (glmfit + candidate-set BH + significance reappearance ≥ 0.6)	0.58	0.26	Parallel dual-channel C end-to-end 0.62; single-test calibration loose (12% vs. 5%)
v11 partitioned-variable residualization	0.34	0.23	False positives not reduced; substantial sensitivity loss
Split-point location stability gate (± 0.2 bandwidth)	0.24	0.10	False positives controlled but sensitivity halved
Pure test (BH only, 20-replicate quick test)	0.95	0.05–0.15	Highest sensitivity but uncontrollable false positives; not included in the final configuration

3.9 Real-data case analyses

3.9.1 CRRT cohort

Among all 595 candidate variable pairs from the 31 variables, the SHAP interaction strength top-5 were $\text{Age} \times P_{24h}$, $P_{24h} \times \text{CRRT_dose}$, $P_{24h} \times \text{SOFA_CRRT_0h}$, $\text{alb_0h} \times \text{SOFA_CRRT_0h}$, and $P_{24h} \times \text{CCI}$ (Table 10). At the confirmation stage, only $P_{24h} \times \text{CCI}$ passed the Wald test ($p = 0.0011$) and BH correction ($p = 0.0053$) and remained robust under rcs main effects ($p = 0.0015$), but its bootstrap reproducibility was 0.84, below the 0.9 threshold, so it was assigned to Level 3 "preliminary confirmation." $\text{Age} \times P_{24h}$, the pair with the highest SHAP strength, did not pass BH correction ($p = 0.088$) and had a reproducibility of only 0.53, placing it at Level 2 "preliminary evidence." The remaining three pairs had Wald $p > 0.3$ and were only "candidate indications." This case demonstrates the process by which the framework converges from hundreds of candidate pairs to a single candidate and assigns evidence grades on real clinical data. P_{24h} (24-hour phosphate) $\times$ CCI (Charlson comorbidity index) was identified as the only interaction candidate with statistical support, providing a testable interaction hypothesis for subsequent research. At the same time, the candidate pair with the highest SHAP strength ranking ($\text{Age} \times P_{24h}$) was not confirmed, consistent with the simulation warning that "high-scoring SHAP candidates may be spurious interactions." Following the mechanism-interpretation positioning, MOB mechanism refinement was further applied to $P_{24h} \times \text{CCI}$, which had passed BH correction (binary approximation; Table 10 note). In the bidirectional focused tests, the direction partitioning on CCI

focusing on the P_{24h} coefficient was significant ($p = 0.017$) with a bootstrap significance-reappearance rate of 0.80 and a split point at CCI = 2. The segment effects showed P_{24h} OR = 1.42 ($p < 0.001$) in the $CCI \leq 2$ group ($n = 540$) and OR = 1.27 ($p = 0.009$) in the $CCI > 2$ group ($n = 604$), i.e., the mortality-risk association of elevated phosphate was stronger in the lower comorbidity-burden subgroup. The reverse direction (P_{24h} partitioning focusing on the CCI coefficient) was not significant ($p = 0.33$), supporting CCI as the modifying variable. This refinement converts the statistically confirmed interaction pair into a testable mechanism hypothesis—"comorbidity burden strata modify the phosphate effect."

To make this result easier to understand, an intuitive mechanistic reading is given here. MOB differs from conventional interaction tests in the question it asks. A conventional product-term test assumes the interaction effect is the same in all patients, whereas MOB allows the effect to vary across patient groups defined by a feature and automatically searches for the split point that maximizes between-group effect differences. The bidirectional test answers "who modifies whom": CCI determines the strength of the phosphate effect (direction B significant), rather than phosphate determining the CCI effect (direction A not significant); that is, the interaction structure is unidirectional. The split point $CCI = 2$ was chosen by the data, not a prespecified threshold. The segment effects show that each 1-unit increase in phosphate multiplies the mortality risk by 1.42 in the low-comorbidity group ($CCI \leq 2$) and by 1.27 in the high-comorbidity group ($CCI > 2$); neither confidence interval includes 1, so phosphate is associated with mortality risk in both groups, but the strength of the association differs. The biological intuition can be understood as a "ceiling effect": patients with high comorbidity already have a high baseline mortality risk, so the incremental harm of elevated phosphate is relatively limited, whereas in relatively healthy patients elevated phosphate is a more prominent danger signal warranting clinical attention. The bootstrap reappearance rate of 0.80 indicates that this stratified structure remains stable under resampling and is not an artifact of a single sample. It should be noted that this reading belongs to the mechanism-interpretation dimension rather than final confirmation: the cut point was obtained by exploratory search, the test p values were not corrected for global multiple testing, and the outcome was approximated by a binary model, so the conclusion needs to be validated in an independent cohort.

Table 10

Confirmation results for the SHAP top-5 candidate pairs in the CRRT cohort ($n = 1144$, 821 deaths within 90 days).

Candidate pair	SHAP strength	Wald p	r _{cs} p	BH p	Bootstrap reproducibility	Evidence grade
Age $\times P_{24h}$	0.0839	0.035	0.192	0.088	0.53	Level 2 – Preliminary evidence
$P_{24h} \times \text{CRRT_dose}$	0.0546	0.829	0.793	0.829	—	Candidate indication
$P_{24h} \times \text{SOFA_CRRT_0h}$	0.0501	0.348	0.427	0.476	—	Candidate indication
alb_0h $\times \text{SOFA_CRRT_0h}$	0.0497	0.381	0.650	0.476	—	Candidate indication
$P_{24h} \times \text{CCI}$	0.0471	0.001	0.002	0.005	0.84	Level 3 – Preliminary confirmation

Note. MOB mechanism refinement for $P_{24h} \times \text{CCI}$ (mechanism-interpretation dimension, binary approximation). Direction with CCI partitioning focusing on the P_{24h} coefficient: $p = 0.017$, bootstrap significance-reappearance rate 0.80, split point CCI = 2; segment effects: P_{24h} OR = 1.42 ($p < 0.001$) in the $\text{CCI} \leq 2$ group and OR = 1.27 ($p = 0.009$) in the $\text{CCI} > 2$ group. Reverse direction (P_{24h} partitioning focusing on the CCI coefficient): $p = 0.33$.

3.9.2 GMDR data

The gold-standard pair ITGA2rs4865756 $\times$ HABP2rs7923349 ranked 9th of 21 in SHAP interaction strength and did not enter the top-5 candidate set. The linear dose product-term Wald test gave $p = 0.093$, and the factor-by-factor (3 $\times$ 3 genotype combination) likelihood-ratio test gave $p = 0.412$; neither statistical metric was significant (Table 11). By contrast, in the single-SNP main effects, the dose effect of HABP2rs7923349 was highly significant ($p = 1.2 \times 10^{-4}$), and the prevalence of the TT genotype (23.9%) was significantly higher than that of GG (10.8%), consistent with the original paper's report of TT as an independent risk factor. To clarify the nature of this "known interaction," a replication test was performed following the original paper's high-risk interacting-genotype analysis structure. Genotype combinations with prevalence above the overall mean were defined as high-risk combinations and entered a logistic model together with the HABP2rs7923349 TT main effect; the replicated high-risk-combination OR was 1.44 (95% CI 1.11–1.86, $P = 0.005$), almost identical to the reported OR = 1.42 ($P = 0.008$), confirming consistency between the data and the testing framework. However, dissecting the source of significance reveals the following. In the cell-by-cell tests, only combinations containing the TT allele (AG $\times$ TT, GG $\times$ TT) were significant, and all significant cells contained TT; TT carriers fell into the high-risk group in 100% of cases; and the key cell AA $\times$ TT, which would support a modifying-effect difference, had only 6 observations. The literature-reported interaction is thus a weak signal driven by "post hoc aggregation of TT-containing high-risk combinations," with insufficient evidence under global statistical testing—a product of a weak signal and a lenient testing paradigm (cell splitting, post hoc selection, and lack of global correction). Here the framework plays the role of an "evidence re-examiner": even when the literature claims an interaction, the framework gives a negative conclusion under a uniform standard. Consistent with the mechanism-interpretation positioning of this study (MOB mechanism refinement is applied only to interactions that have passed BH correction), the gold-standard pair in this case did not pass confirmation and had no mechanism-refinement object, so MOB analysis was not performed.

Table 11

Validation results for the gold-standard interaction pair ITGA2rs4865756×HABP2rs7923349 in the GMDR data ($n = 2377, 295$ carotid stenosis cases).

Metric	Value
SHAP interaction strength rank (/21)	9th
Linear dose product-term Wald p	0.093
Factor-by-factor likelihood-ratio p (4 df)	0.412
HABP2rs7923349 dose main-effect p	1.2×10^{-4}
TT vs. GG genotype prevalence	23.9% vs. 10.8%
Original paper GMDR permutation P	0.010
High-risk-combination replication (this study, with TT main effect)	OR 1.44 (1.11–1.86), $P = 0.005$ (paper 1.42, $P = 0.008$)
Significant high-risk cells (cell-by-cell vs. GG×GG)	Only TT-containing combinations (AG×TT, GG×TT); key cell AA×TT $n = 6$

4. Discussion

4.1 Rationale of the two-stage design and evidence grading

The suitability evaluation yields the following conclusions. At the screening stage, SHAP interaction and Friedman H have strong candidate-convergence capability (98%/65% detection of nonlinear interactions in scenario C, far exceeding the 85% of the linear product-term test) but lack significance control (false positives 64%/30% in scenario A), making them suited to the "candidate shrinkage" task. The within-family comparison at the screening stage (v12) further showed that statistical-test ranking (SIS type) has slightly lower coverage than model attribution in the threshold-type nonlinear scenario (90% vs. 99%) and, being informationally homogeneous with the confirmation stage, drives the candidate-set false-confirmation rate up to 96% in the no-interaction scenario (vs. 58% for SHAP); hence the screener should select a ranking signal heterogeneous from the confirmer. At the confirmation stage, the product-term Wald test and GAM provide controlled significance evidence (false positives 6%/7% in scenario A, at the nominal level) but have insufficient sensitivity to nonlinear interactions (detection 85%/81% in scenario C), making them suited to the "true/false judgment" task. MOB has potential at both ends, but as an independent evidence channel its false positives are uncontrolled (26% false alarms in scenario A in v10), and its proper position is mechanism refinement after confirmation. This division of "heavy coverage at the screening stage and heavy control at the confirmation stage" is the design basis of the two-stage framework and constitutes a direct guide for method choice: select the screener (tree-model attribution or marginal statistic) and confirmer (parametric or semiparametric test) according to the data structure, and retain FDR control and resampling robustness at the confirmation stage.

The core finding of this study is that the necessity of the "screen–confirm" division of labor is confirmed by simulation. SHAP interaction alone had false positives as high as 64% under collinearity without interaction, statistical tests alone detected under 85% under threshold-type nonlinear interactions, and the two-stage method reduced the former by an order of magnitude (end-to-end false positives 5% in scenario A, further to 7% false confirmation with zero end-to-end under the v9 configuration) while maintaining the latter at 83% (v3). This result is consistent with the theoretical property that SHAP interaction values are a game-theoretic decomposition carrying

no significance information (Lundberg et al., 2018), and echoes REPID's conclusion that existing global interaction-detection methods have shortcomings (Herbinger et al., 2022). GGInt-XGBoost provides p values for tree-model interaction structures directly through permutation tests (Guo et al., 2021); the v8 permutation-calibration experiment of this study shows that this idea has limited filtering power for collinearity-transmitted spurious interactions, indicating that "substituting permutation tests for statistical confirmation" is not a sufficient condition, and the confirmation stage still needs a classical framework with FDR control and resampling robustness. The resulting four-level evidence grading reports "candidate indication—preliminary evidence—preliminary confirmation—robust confirmation" separately, with mechanism interpretation presented as an independent post-confirmation dimension, avoiding the misreading of screening scores as significance—the most important usage norm of the framework in practice. The systematic validation of MOB as an independent evidence channel (v10) delineates the boundary of this grading. MOB captured 58% of threshold-type interactions (62% parallel end-to-end) but had a 26% false-alarm rate on spurious pairs in the no-interaction scenario with loose single-test p -value calibration; the two control strategies—residualization and a split-point location gate—ended in ineffectiveness and sensitivity loss, respectively, showing that its high sensitivity comes at the price of uncontrolled specificity and that it should not serve as independent confirmation evidence. Its proper position is a mechanism-refinement tool after statistical confirmation (reporting split points and segment effects), and this boundary is itself an empirical conclusion of this study, demonstrated on real data: the $P_{24h} \times CCI$ interaction in the CRRT case yielded a CCI = 2 cut point and segment effect sizes through MOB mechanism refinement, converted into the testable mechanism hypothesis "comorbidity burden strata modify the phosphate effect." The real-data case analyses further corroborate the operability of the framework. In the CRRT cohort, the framework converged from 595 candidate pairs to the single Level 3 "preliminary confirmation" candidate $P_{24h} \times CCI$ without inflating the evidence to Level 4 "robust confirmation"; in the GMDR data, it rejected a published but unreproducible interaction conclusion, avoiding the false-positive propagation of "literature endorsement." The two cases correspond to the two facets of the framework, "screening out candidates" and "rejecting spurious interactions," showing that evidence grading constrains not only the strength but also the direction of conclusions.

4.2 Comparison with related work

Methodologically, the framework sits at the intersection of the SIS lineage and machine-learning interpretability tools. SIS and its extensions achieve ultrahigh-dimensional reduction through marginal correlations or marginal likelihoods (Fan & Lv, 2008; Fan et al., 2009), with TS-GSIS and EPISIS extending this to genome-wide interaction settings (Fang et al., 2017; Hibar et al., 2015). Their two-stage idea of "screen first, select later" is consistent with the screen—confirm structure of this study; the difference is that the screener here is based on in-model SHAP interaction strength rather than marginal quantities, allowing it to capture interaction signals without marginal main effects. Correspondingly, the top- k evaluation (v4) of this study quantifies the trade-off between screening width and candidate contamination, which can serve as a simulation reference for candidate-set size selection of SIS-type methods in interaction settings. Reluctant Interaction Modeling advocates prioritizing main effects over interactions (Yu et al., 2019), which is in spirit consistent with the "all main effects must enter the model" discipline of this study, but it handles interactions at the fitting stage whereas this study handles them at the testing stage, and the problems addressed differ (predictive modeling vs. interaction inference). SHAP-XGB reported that SHAP interactions outperform conventional tests in QTL interaction mapping (Ishibashi & Onogi, 2024), in a setting where main effects and interactions coexist with good variable independence; the 64% false positives revealed in scenario A of this study occur in a strong-collinearity, no-interaction setting. The two are complementary, showing that the applicability of SHAP interactions depends on the data structure, and the collinearity robustness grid (v6) of this study provides a quantitative characterization of this boundary. The two negative results of this study

(residualization preprocessing, permutation calibration) also constitute a methodological contribution: they rule out two intuitively plausible but actually harmful or inefficient paths, saving exploration cost for subsequent method selection. The real-data case analyses further reveal a contrast with traditional nonparametric interaction-detection methods. GMDR-type methods (multifactor dimensionality reduction) select the best interaction model by cross-validation consistency and sign tests and define high-risk genotype combinations by cell means; their permutation test insufficiently corrects for the selection bias of the "best model," and the paradigm of "post hoc selection of high-prevalence combinations followed by aggregated testing" lacks global multiple-comparison control. Although the GMDR-reported ITGA2rs4865756×HABP2rs7923349 interaction in this study could be reproduced under the original structure (high-risk-combination OR = 1.44), the global statistical tests (dose product $p = 0.093$, factor-by-factor $p = 0.412$) could not confirm it; all significant high-risk cells contained the TT allele and the key cell had only 6 observations, suggesting that this positive result is more likely a product of a weak signal and a lenient testing paradigm than a robust interaction. Positive results from nonparametric interaction detection need independent statistical-test re-examination—precisely the role played by the confirmation stage of the framework here.

4.3 Applicability and limitations

Based on the robustness grid, the recommended usage conditions of the framework are sample size $n \geq 500$ (≥ 250 events), event rate $\geq 30\%$, collinearity strength $\rho \leq 0.7$, and binary or survival outcomes. At $n < 500$ or event rates $< 10\%$, detection of threshold-type interactions is unreliable (detection $< 40\%$), and negative results should be interpreted with caution. The main limitations are fivefold. (1) The false-confirmation rate in the strong-linear-interaction scenario (scenario B) remains elevated under some configurations—39% under the final configuration (v9 wald09) and up to 72% at the Cox-outcome grid point of the robustness grid—an inherent challenge of "indistinguishability between collinear spurious interactions and true interactions" that requires multiple evidence sources or domain priors. (2) The simulation covers linear and single-threshold interaction forms but not more complex multi-layer or higher-order interaction structures. (3) Screening uses XGBoost as the only base model; interaction decompositions of other tree ensembles (random forest, LightGBM) and deep models were not evaluated. (4) The dimensionality $p = 10$ is lower than the ultrahigh-dimensional settings of real omics data, and computational strategies for the explosion of candidate pairs at high dimensionality, such as RIM's blockwise solving idea, were not included. (5) The real-data validation has inherent limitations. The CRRT case lacks an external gold standard, and $P_{24h} \times CCI$ reached only Level 3 "preliminary confirmation" evidence, requiring validation in an independent cohort. The GMDR public data do not include the covariates adjusted in the original paper (age, hypertension, diabetes, smoking, lipids), and covariate adjustment could affect the interaction-test conclusions; in this case the SHAP screening stage also failed to rank the gold-standard pair into the top-5, indicating limited screening sensitivity of the method for weak interactions, and the case is essentially a negative result of "rejecting a reported interaction" rather than a positive confirmation of detection capability. These directions are left to future work.

5. Conclusion

This study demonstrates through systematic simulation evaluation that, around the two-stage demand of "screening plus confirmation," interaction-analysis methods exhibit a clear suitability division: tree-model attribution methods (SHAP) and marginal statistics (Friedman H) suit the screening stage; parametric/semiparametric tests (Wald, GAM) suit the confirmation stage; and model-based recursive partitioning (MOB) is limited to mechanism refinement after confirmation. The resulting two-stage detection pipeline (SHAP interaction top- k screening with $k = 5$, product-term Wald test, BH-FDR correction, and bootstrap reproducibility ≥ 0.9 robust confirmation) balances

the trade-off between a high false-positive rate in screening and insufficient sensitivity in testing and organizes its output through a four-level evidence-grading framework. The mechanism-interpretation channel (MOB) was shown to have high sensitivity but uncontrolled false positives and is used only as a mechanism-refinement tool for confirmed interactions, not in the end-to-end decision. The method performs robustly on binary and survival data with sample size $n \geq 500$, event rate $\geq 30\%$, and collinearity $\rho \leq 0.7$; collinearity residualization, permutation noise calibration, and the MOB independent evidence channel were all falsified as ineffective or inefficient paths. In the real-data case analyses, the method screened out $P_{24h} \times CCI$ in the CRRT cohort and assigned it to Level 3 "preliminary confirmation," and rejected the unreproducible reported interaction in the GMDR data, preliminarily corroborating its dual capability of "screening out candidates" and "rejecting spurious interactions." The study provides a simulation-evidence-chain-based method-suitability reference and an operational usage boundary for interaction detection in medical and epidemiological settings.

Declarations

Ethics approval and consent to participate

Not applicable. This study analyzed publicly available, de-identified datasets and did not involve human participants, human material, or the collection of new data.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Funding

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

Authors' contributions

Yue-peng Liu conceived and designed the study, developed the simulation framework, performed the data analyses, interpreted the results, and wrote and revised the manuscript. The author read and approved the final manuscript.

Availability of data and materials

The datasets analyzed in this study are publicly available from the Dryad Digital Repository: the CRRT cohort data (Jung et al., 2019, <https://doi.org/10.5061/dryad.6v0j9>) and the GMDR carotid-stenosis data (Liu, 2022, <https://doi.org/10.5061/dryad.k98sf7m9r>). The R code used for the simulations and analyses is available from the corresponding author on reasonable request.

References

1. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. *J Roy Stat Soc B* 70(5):849–911
2. Fan J, Samworth R, Wu Y (2009) Ultrahigh dimensional feature selection: Beyond the linear model. *J Mach Learn Res* 10:2013–2038
3. Fang Y-H, Wang J-H, Hsiung CA (2017) TS-GSIS: A high-dimensional grouped variable selection approach for detection of whole-genome SNP–SNP interactions. *Bioinformatics* 33(22):3595–3602
4. Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Annals Appl Stat* 2(3):916–954
5. Guo Y, Wu C, Yuan Z et al (2021) Gene-based testing of interactions using XGBoost in genome-wide association studies. *Front Cell Dev Biology* 9:801113
6. Herbringer J, Bischl B, Casalicchio G (2022) REPID: Regional effect plots with implicit interaction detection. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022)*. *arXiv:2202.07254*
7. Hibar DP, Stein JL, Jahanshad N et al (2015) Genome-wide interaction analysis reveals replicated epistatic effects on brain structure. *Neurobiol Aging* 36:S151–S158. <https://doi.org/10.1016/j.neurobiolaging.2014.02.033>
8. Ishibashi T, Onogi A (2024) Applying gradient tree boosting to QTL mapping with Shapley additive explanations. *bioRxiv*
9. Jung S-Y, Kwon J, Park S et al (2018) Phosphate is a potential biomarker of disease severity and predicts adverse outcomes in acute kidney injury patients undergoing continuous renal replacement therapy. *PLoS ONE*, 13(2), e0191290
10. Jung S-Y et al (2019) Data from: Phosphate is a potential biomarker of disease severity and predicts adverse outcomes in acute kidney injury patients undergoing continuous renal replacement therapy. Dryad Digit Repository. <https://doi.org/10.5061/dryad.6v0j9>
11. Liu L (2022) Data from: The associations of gene–gene interactions among the 6 variants in 3 genes related to inflammation and endothelial function with carotid stenosis were performed by the GMDR approach. Dryad Digit Repository. <https://doi.org/10.5061/dryad.k98sf7m9r>
12. Liu L, Yi X, Luo H, Yu M (2023) Inflammation and endothelial function relevant genetic polymorphisms in carotid stenosis in southwestern China. *Front Neurol* 13:1076898
13. Lundberg SM, Erion GG, Lee S-I (2018) Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*
14. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2(1):56–67
15. Pan W, Wang X, Xiao W, Zhu H (2019) A generic sure independence screening procedure. *J Am Stat Assoc* 114(526):928–937. <https://doi.org/10.1080/01621459.2018.1462709>
16. Yu G, Bien J, Tibshirani R (2019) Reluctant interaction modeling. *arXiv preprint arXiv:1907.08414*
17. Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. *J Comput Graphical Stat* 17(2):492–514