Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identifying Sociodemographic, Clinical, and Behavioral Determinants of Self- Reported Kidney Stone History in US Adults: A Population-Based Comparative Analysis of Logistic Regression, Random Forest, and XGBoost Using NHANES 2007–2018

Chengbin Li[1]
Xi'an Daxing Hospital
Hao Wu[1]
Xi'an Daxing Hospital
Dawei Zhao[1]
Xi'an Daxing Hospital
Yanzhu Wang[1]
wangyanzhu1970@126.com

Xi'an Daxing Hospital

Research Article

Keywords: kidney stones, nephrolithiasis, NHANES, machine learning, random forest, XGBoost, SHAP, logistic regression, nomogram

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Identifying Sociodemographic, Clinical, and Behavioral Determinants of Self-Reported Kidney Stone History in US Adults: A Population-Based Comparative Analysis of Logistic Regression, Random Forest, and XGBoost Using NHANES 2007–2018

Chengbin Li[1], Hao Wu[1], Dawei Zhao[1], Yanzhu Wang[1]*

[1] Department of Urology, Xi'an Daxing Hospital, Laodong North Road, Lianhu District, Xi'an, Shaanxi Province, China

* Corresponding author: Yanzhu Wang

Department of Urology, Xi'an Daxing Hospital, Laodong North Road, Lianhu District, Xi'an, Shaanxi Province, China

Email: wangyanzhu1970@126.com | ORCID: 0009-0005-3710-8511

Author emails and ORCIDs:

Chengbin Li: leechengbin0404@gmail.com | ORCID: 0009-0002-3326-4654

Hao Wu: 312388696@qq.com | ORCID: 0009-0000-1603-8239

Dawei Zhao: zhaodawei3172010@163.com | ORCID: 0009-0004-2299-2060

Abstract

Background: Logistic regression remains conventional for identifying correlates of self-reported kidney stone history, but multidimensional comparisons with machine learning—integrating discrimination, calibration, and clinical utility—remain scarce. This study aimed to identify determinants of self-reported kidney stone history among US adults and to compare logistic regression, random forest, and XGBoost.

Methods: We conducted a pooled cross-sectional analysis of six consecutive NHANES cycles (2007–2018), including adults aged ≥20 years with complete data on kidney stone history and 16 prespecified candidate predictors. LASSO logistic regression with 10-fold cross-validation selected variables in a 70% training subset.

Logistic regression, random forest, and XGBoost were trained on the identical predictor set and evaluated on the held-out 30% test set by discrimination (AUC with DeLong's test), calibration (intercept and slope), and clinical utility (decision curve analysis). Survey-weighted sensitivity analysis, collinearity assessment, nomogram, and SHAP analyses.

Results: Of 25,560 adults, 2,462 (9.6%; survey-weighted 9.99%) reported a lifetime history of kidney stones. LASSO retained 14 of 16 candidates. Eight variables were independently associated with kidney stone history: older age, male sex, lower BMI (inverse), higher waist circumference, elevated blood urea nitrogen, hypertension, diabetes mellitus, and current smoking (all 95% CIs excluding unity). BMI and waist circumference showed substantial collinearity (VIFs 7.69 and 7.81). Logistic regression achieved an AUC of 0.644, comparable to XGBoost (0.641; DeLong P = 0.591) and superior to random forest (0.617; P = 0.0005), with the strongest calibration (intercept −0.147; slope 0.892), followed by XGBoost (slope 0.771) and random forest (slope 0.569). On decision curve analysis, random forest's net benefit fell below the treat-none line within the 15–35% threshold range, whereas logistic regression and XGBoost remained stable. SHAP analysis confirmed age, waist circumference, and blood urea nitrogen as the top contributors; survey-weighted results were consistent.

Conclusions: Older age, male sex, adiposity-related traits, elevated blood urea nitrogen, hypertension, diabetes mellitus, and current smoking were independently associated with prevalent kidney stone history. Logistic regression matched XGBoost in discrimination while showing the best calibration, stable clinical net benefit, and full interpretability. Given the cross-sectional, self-reported design, these findings reflect correlates rather than causal predictors and warrant external validation.

Keywords: kidney stones; nephrolithiasis; NHANES; machine learning; random forest; XGBoost; SHAP; logistic regression; nomogram

Introduction

Nephrolithiasis is a prevalent and recurrent urological disorder, affecting an estimated 10% of US adults over their lifetime [1, 2], with recurrence rates approaching 50% within five years following an initial episode.[12] ts pathophysiology is multifactorial—driven by interrelated metabolic, dietary, anthropometric, and behavioral determinants—and its rising population-level burden parallels the epidemiologic trends of obesity, hypertension, and type 2 diabetes mellitus.

Epidemiologic studies of kidney stone risk have historically relied on multivariable logistic regression.[14,15] While this approach offers strong interpretability and well-established inference frameworks, it presupposes approximately linear logit–predictor relationships and limited capacity to detect unanticipated non-linearities or higher-order interactions—particularly among correlated metabolic and anthropometric variables (e.g., BMI and waist circumference). In contrast, machine learning (ML) algorithms—including ensemble methods such as random forest and gradient-boosted decision trees (e.g., XGBoost)—are increasingly adopted in clinical and epidemiologic research due to their flexibility in modeling complex, non-linear patterns and implicit interaction effects without requiring a priori specification. Yet rigorous, multidimensional model comparisons—systematically evaluating discrimination alongside calibration and clinical decision utility—remain underrepresented in the nephrolithiasis literature, especially within large, nationally representative, longitudinal survey data spanning more than a decade.[17]

Critically, kidney stone history in cross-sectional surveys like NHANES is a self-reported, lifetime prevalence outcome—not incident disease[13]. As such, statistical models fitted to these data are appropriately interpreted as tools for identifying and quantifying correlates of established stone disease, rather than prospective predictors of future stone formation[20]. Several candidate predictors—including blood urea nitrogen, hypertension, and diabetes—are contemporaneously measured with the outcome; thus, observed associations may reflect sequelae or comorbid manifestations of chronic stone-related pathophysiology (e.g., renal impairment, metabolic syndrome) rather than strictly antecedent risk factors[21]. This conceptual distinction between association and prediction—and its implications for causal inference—is explicitly acknowledged throughout the analysis and interpretation.

The National Health and Nutrition Examination Survey (NHANES) provides a rigorously designed, nationally representative, publicly accessible dataset containing comprehensive demographic, anthropometric, biochemical, dietary, and self-reported health information. Its complex sampling structure, standardized protocols, and longitudinal coverage across multiple cycles (2007–2018) make it uniquely suited for developing,

internally validating, and survey-weighting statistical and machine learning models of kidney stone history at the population level.

Therefore, this study aimed to: (1) identify sociodemographic, clinical, and behavioral determinants independently associated with self-reported kidney stone history among US adults using pooled NHANES data (2007–2018); (2) conduct a head-to-head, multimetric evaluation of logistic regression, random forest, and XGBoost—including discrimination (AUC), calibration (intercept/slope), and clinical net benefit (decision curve analysis); (3) apply SHAP analysis to elucidate feature importance and contribution patterns in the top-performing tree-based model; and (4) develop a validated, clinically deployable nomogram from the final logistic regression model to support individualized estimation of lifetime kidney stone probability.

Methods

Study design, data source, and reporting

This cross-sectional study used data from six consecutive NHANES cycles (2007–2008 through 2017–2018), a nationally representative survey conducted by the National Center for Health Statistics (NCHS), U.S. Centers for Disease Control and Prevention (CDC). NHANES employs a complex, multistage probability sampling design and combines household interviews, physical examinations, and laboratory testing. All NHANES protocols were approved by the NCHS Research Ethics Review Board, and written informed consent was obtained from all participants; as this study used publicly available, de-identified data, additional institutional review board approval was not required. This study was designed and reported with reference to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement, adapted for a cross-sectional, correlate-identification context [3].

Study population

Among 59,842 examined participants across the six cycles, 34,770 were aged ≥20 years. Of these, 34,679 (99.7%) provided a definitive (yes/no) response regarding kidney stone history. After further excluding 9,119 participants with missing data on one or more of 16 candidate predictor variables, 25,560 participants were included in the final analytic cohort (Figure 1).

Outcome variable

The outcome, history of kidney stones, was defined by an affirmative response to "Have you ever had kidney stones?" (variable KIQ026) in the Kidney Conditions–Urology questionnaire. This is a self-reported, lifetime (ever/never) outcome rather than an incident or time-to-event measure.

Candidate predictor variables and variable definitions

Sixteen candidate predictors were pre-specified based on established pathophysiologic and epidemiologic evidence: age (RIDAGEYR) and sex (RIAGENDR); anthropometric measures—body mass index (BMXBMI) and waist circumference (BMXWAIST); laboratory biomarkers—serum uric acid (LBXSUA), serum calcium (LBXSCA), blood urea nitrogen (LBXSBU), and non-fasting serum glucose (LBXSGL); self-reported physician-diagnosed hypertension and diabetes; current smoking status; poverty–income ratio (INDFMPIR); and dietary intakes—including sodium, calcium, potassium, and total plain water—estimated from a single 24-hour dietary recall.

Key variable definitions were rigorously operationalized to ensure reproducibility and conceptual fidelity:

• Hypertension and diabetes were defined by affirmative responses to standardized NHANES questions ("Ever told you had high blood pressure?" [BPQ020] and "Ever told you had diabetes?" [DIQ010]), reflecting clinical diagnosis rather than physiological thresholds. This aligns with population-level surveillance practices and avoids misclassification due to transient or unmeasured blood pressure or glycemic variability.

• Serum glucose (LBXSGL) represents a non-fasting (random) measurement obtained as part of the standard biochemistry profile for all examined participants—distinct from fasting plasma glucose (LBXGLU), which is collected only in the morning fasting subsample. Given its broader availability and relevance to metabolic burden, LBXSGL was included a priori but excluded during LASSO variable selection due to lack of stability and marginal contribution.

• Current smoking status was derived hierarchically: participants reporting lifetime use of $\geqslant$100 cigarettes (SMQ020 = "Yes") *and* current smoking frequency of "Every day" or "Some days" (SMQ040) were classified as current smokers; those reporting $\geqslant$100 cigarettes but no current use (SMQ040 = "Not at all") were classified as former smokers; and those reporting <100 cigarettes (SMQ020 = "No") were classified as never smokers. The binary current smoker (yes/no) variable was used as the primary exposure in all models to reflect active tobacco exposure relevant to stone pathogenesis.

Statistical analysis

Continuous variables are presented as mean (SD) and were compared using Student's *t*-test; categorical variables are presented as frequency (percentage) and compared using the chi-square test or Fisher's exact test where expected cell counts were <5. The primary analysis used unweighted data to maximize statistical power and model stability; a survey-weighted sensitivity analysis—accounting for NHANES complex sampling design—is reported separately to ensure nationally representative inference.

With 2,462 outcome events and a maximum of 16 candidate predictors, the events-per-variable (EPV) ratio exceeded 150, well above commonly cited minimum thresholds for stable regression estimation.

Participants were randomly divided into training (70%, n=17,892) and testing (30%, n=7,668) sets. LASSO regression with 10-fold cross-validation was applied within the training set for variable selection; the lambda.min criterion was used for the primary variable set, and the more parsimonious lambda.1se variable set was evaluated as a sensitivity analysis [25]. Variables retained by LASSO (lambda.min) were entered into a multivariable logistic regression model. All models—logistic regression, random forest, and XGBoost—were trained on the identical LASSO-selected predictor set to ensure methodologically fair comparison.

Multicollinearity among retained predictors was formally assessed using variance inflation factors (VIF); VIF >5 indicated moderate collinearity and informed interpretation of effect estimates. Random forest was implemented with 500 trees and default hyperparameters and XGBoost (200 boosting rounds, maximum depth 4, learning rate 0.05) models were trained on the identical predictor set for direct comparison [26, 32].

Model discrimination was assessed using the AUC in the independent test set, with pairwise comparisons between models performed using DeLong's test [4]. Calibration was assessed in the independent test set for all three models by regressing the observed outcome on the logit of the predicted probability, yielding a calibration intercept (calibration-in-the-large) and slope (ideal values of 0 and 1, respectively); a bootstrap-corrected (1,000 resamples) internal calibration curve was additionally produced for the logistic model. Clinical utility was assessed using decision curve analysis (DCA) [5] comparing net benefit of all three models against "treat-all" and "treat-none" strategies across threshold probabilities. A nomogram was constructed from the final logistic model to enable individualized estimation of the probability of kidney stone history [27].

SHAP (SHapley Additive exPlanations) values were computed for the XGBoost model to quantify the contribution of each variable to individual predictions, results were visualized using both SHAP summary (beeswarm) and dependence plots to examine directionality, magnitude, and potential interactions.

The survey-weighted sensitivity analysis was conducted using the R survey package , incorporating six-cycle

combined examination weights (WTMEC2YR/6), masked variance pseudo-strata (SDMVSTRA), and pseudo-primary sampling units (SDMVPSU). This analysis yielded nationally representative estimates of kidney stone prevalence and weighted odds ratios, enabling direct comparison with the primary unweighted findings.

All analyses were performed in R (version 4.6.1) using the following packages: glmnet (v4.1-8), rms (v6.7-1), pROC (v1.18.0), randomForest (v4.7-1.1), xgboost (v1.7.1), shapviz (v0.2.2), car (v3.1-2), survey (v4.2-1), and dcurves (v0.4.0). Statistical significance was defined as two-sided $P < 0.05$ or 95% confidence intervals excluding the null value (odds ratio = 1).

This study adheres to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement, with adaptations for cross-sectional prevalence modeling—including explicit acknowledgment that associations reflect correlates of established disease rather than prospective risk predictors. A completed TRIPOD checklist is provided in Additional file 1.

Results

Study population and flow

Figure 1 shows the participant flow. Among 25,560 adults in the final analytic cohort, 2,462 (9.6%) reported a history of kidney stones; the survey-weighted prevalence estimate was 9.99% (SE 0.27%).

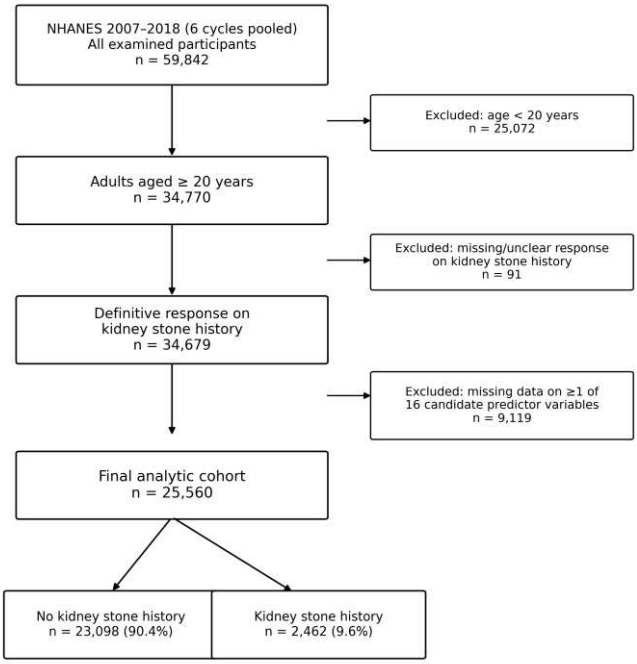


Figure 1. Participant flow diagram.

Baseline characteristics

Compared with participants without a stone history, those reporting a stone history were older, more frequently male, and had higher BMI, waist circumference, serum uric acid, and blood urea nitrogen (all $P<0.001$), with modestly lower serum calcium ($P=0.019$). Hypertension and diabetes were more prevalent among those with a stone history (both $P<0.001$). Notably, the unadjusted prevalence of current smoking was nearly identical between groups (20.4% vs. 20.4%, $P=1.000$); as detailed in the Discussion, this crude null association contrasts with a significant adjusted association, illustrating the influence of confounding by age and other factors. Dietary sodium intake and total water intake did not differ significantly between groups, while dietary calcium and potassium intake were modestly lower in the stone-history group ($P=0.006$ and $P=0.037$, respectively). Full baseline characteristics are shown in Table 1.

Table 1. Baseline characteristics stratified by kidney stone history

Characteristic	No stone history (n=23,098)	Stone history (n=2,462)	P value
Age, years, mean (SD)	48.49 (17.49)	55.36 (16.09)	<0.001
Male sex, n (%)	11,113 (48.1)	1,388 (56.4)	<0.001
BMI, kg/m², mean (SD)	29.09 (6.83)	30.50 (6.79)	<0.001

Waist circumference, cm, mean (SD)	99.15 (16.38)	104.57 (16.21)	<0.001
Serum uric acid, mg/dL, mean (SD)	5.43 (1.43)	5.64 (1.51)	<0.001
Serum calcium, mg/dL, mean (SD)	9.39 (0.36)	9.37 (0.38)	0.019
Blood urea nitrogen, mg/dL, mean (SD)	13.50 (5.66)	15.08 (7.03)	<0.001
Hypertension[a], n (%)	7,863 (34.0)	1,227 (49.8)	<0.001
Diabetes[a], n (%)	2,699 (11.7)	534 (21.7)	<0.001
Current smoking[b], n (%)	4,708 (20.4)	502 (20.4)	1.000
Poverty-income ratio, mean (SD)	2.52 (1.63)	2.50 (1.61)	0.698
Dietary sodium, mg/day, mean (SD)	3489.88 (1852.62)	3461.59 (1885.52)	0.472
Dietary calcium, mg/day, mean (SD)	931.11 (592.02)	896.88 (555.42)	0.006
Dietary potassium, mg/day, mean (SD)	2627.04 (1268.90)	2571.08 (1241.24)	0.037
Total plain water intake, g/day, mean (SD)	1095.00 (1171.08)	1055.31 (1204.74)	0.111

[a]Self-reported physician diagnosis. [b]Current smoking defined by combining lifetime smoking of ≥100 cigarettes (SMQ020) with current smoking frequency (SMQ040); see Methods.

Variable selection

Of 16 candidate predictors entered into the LASSO model, 14 retained non-zero coefficients at the lambda.min penalty, excluding serum glucose and total plain water intake. At the more conservative lambda.1se penalty, only 4 variables were retained (age, waist circumference, hypertension, and diabetes), which achieved a test-set AUC of 0.633—close to the 14-variable model (0.644)—indicating that most of the discriminative information was concentrated in a small subset of variables (Figure 2).

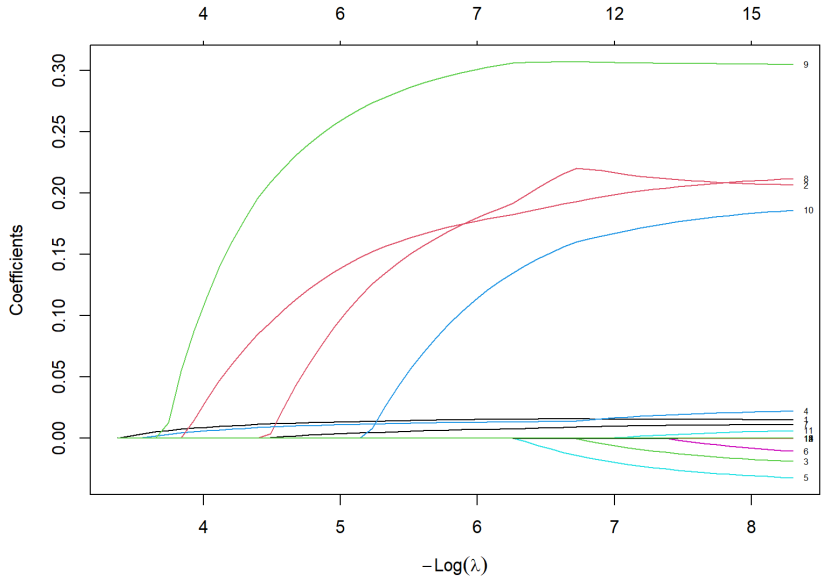


Figure 2A. LASSO coefficient trajectories across the regularization path (variables labeled by index; dashed and dotted vertical lines denote lambda.min and lambda.1se, respectively).

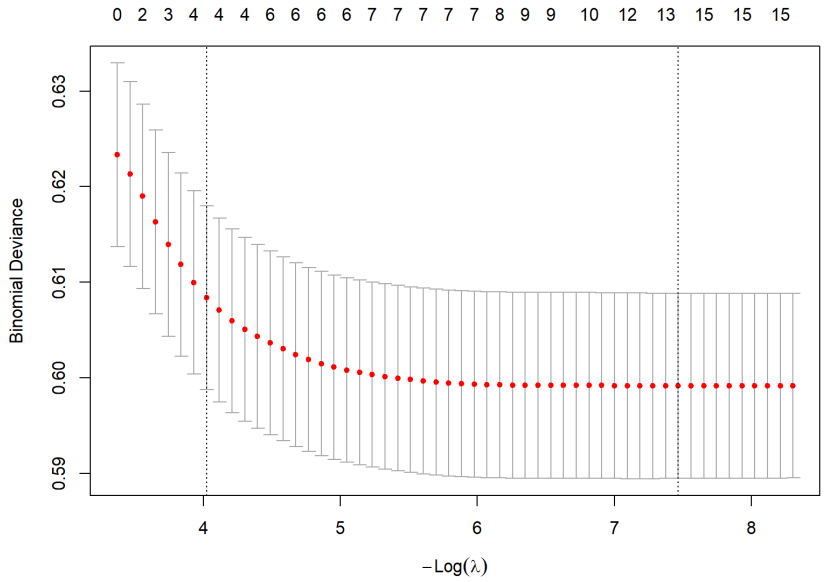


Figure 2B. Ten-fold cross-validation curve showing binomial deviance across log(lambda), with lambda.min and lambda.1se indicated.

Multivariable logistic regression and collinearity assessment

In the multivariable logistic regression model incorporating the 14 LASSO-selected variables, eight variables showed 95% confidence intervals excluding the null value: age (OR 1.015, 95% CI 1.011–1.019), male sex (OR 1.223, 95% CI 1.078–1.387), waist circumference (OR 1.025, 95% CI 1.016–1.034), blood urea nitrogen (OR 1.012, 95% CI 1.003–1.021), hypertension (OR 1.243, 95% CI 1.105–1.397), diabetes (OR 1.357, 95% CI 1.182–1.554), and current smoking (OR 1.212, 95% CI 1.0611.382). BMI showed a modest inverse association (OR 0.976, 95% CI 0.957–0.996). Variance inflation factors were elevated for BMI (7.69) and waist circumference (7.81), indicating substantial collinearity between these two adiposity measures, while all other variables had VIF <2.1; this collinearity likely explains the inverse BMI association observed after mutual adjustment. Serum uric acid, serum calcium, poverty-income ratio, and dietary sodium, calcium, and potassium intake did not reach statistical significance. Full results are presented in Table 2.

Table 2. Multivariable logistic regression results (14 LASSO-selected variables)

Variable	OR	95% CI	Significant (CI excludes 1)
Age (per year)	1.015	1.011–1.019	Yes
Male sex	1.223	1.078–1.387	Yes
BMI (per kg/m²)	0.976	0.957–0.996	Yes
Waist circumference (per cm)	1.025	1.016–1.034	Yes
Serum uric acid (per mg/dL)	0.964	0.925–1.005	No
Serum calcium (per mg/dL)	0.983	0.854–1.130	No
Blood urea nitrogen (per mg/dL)	1.012	1.003–1.021	Yes

Hypertension	1.243	1.105–1.397	Yes
Diabetes	1.357	1.182–1.554	Yes
Current smoking	1.212	1.061–1.382	Yes
Poverty-income ratio	1.009	0.976–1.042	No
Dietary sodium (per mg/day)	1.00002	0.99998–1.00006	No
Dietary calcium (per mg/day)	0.99993	0.99981–1.00005	No
Dietary potassium (per mg/day)	0.99998	0.99992–1.00004	No

Comparison of logistic regression, random forest, and XGBoost

In the independent test set, logistic regression achieved an AUC of 0.644, XGBoost 0.641, and random forest 0.617 (Figure 3). DeLong's test indicated no significant difference between logistic regression and XGBoost (P=0.591), whereas random forest was significantly inferior to both logistic regression (P=0.0005) and XGBoost (P<0.001, from a separate pairwise comparison). Table 3 summarizes discrimination and calibration across the three models.

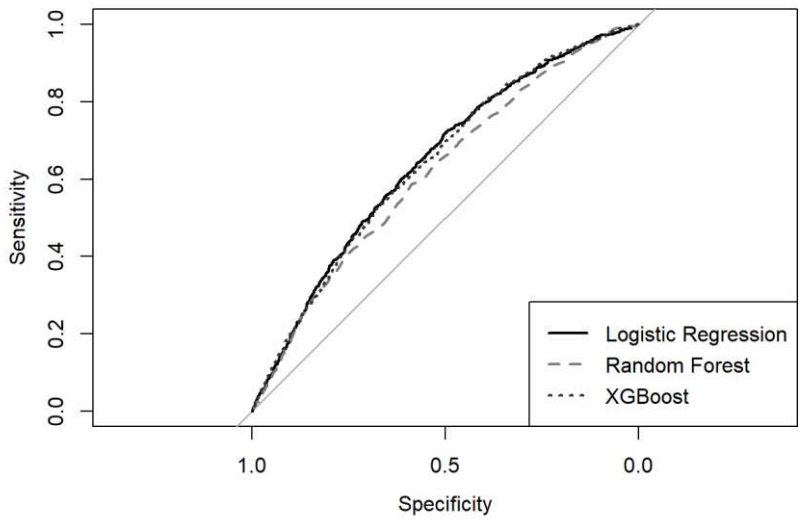


Figure 3. Receiver operating characteristic curves comparing logistic regression, random forest, and XGBoost in the testing set.

Table 3. Discrimination and calibration comparison across models

Model	AUC (test set)	Calibration intercept / slope	vs. Logistic (DeLong P)
Logistic regression	0.644	−0.147 (P=0.342) / 0.892	— (reference)
Random forest	0.617	−0.912 (P<0.001) / 0.569	0.0005
XGBoost	0.641	−0.398 (P=0.004) / 0.771	0.591

Calibration and clinical utility

Independent test-set calibration was optimal for the logistic regression model, with a calibration intercept of −0.147 (P = 0.342; not significantly different from the ideal value of 0) and a slope of 0.892—close to the ideal value of 1—indicating minimal miscalibration in both overall risk level and relative ordering of predicted

probabilities. XGBoost demonstrated moderate calibration (intercept −0.398, P = 0.004; slope 0.771), reflecting modest overestimation of absolute risk and reduced discrimination of risk gradients. In contrast, random forest exhibited poor calibration: its intercept deviated substantially from zero (−0.912, P < 0.001), and its slope was markedly attenuated (0.569), indicating systematic overconfidence—i.e., predicted probabilities were disproportionately inflated relative to observed event rates across the risk spectrum (Figure 4A).

Decision curve analysis (DCA) further revealed divergent clinical utility: logistic regression and XGBoost yielded consistently positive net benefit across nearly the entire clinically relevant threshold range (5–35%), remaining above the "treat-none" reference line—suggesting that using either model to guide risk-informed decisions would confer net clinical advantage over withholding stratification. By contrast, random forest's net benefit declined below the "treat-none" line within the 15–35% threshold probability range (Figure 4B), implying that, at these intermediate risk thresholds, clinical decisions based on its predictions may yield lower net benefit—and potentially greater harm—than applying a uniform management strategy without risk assessment.

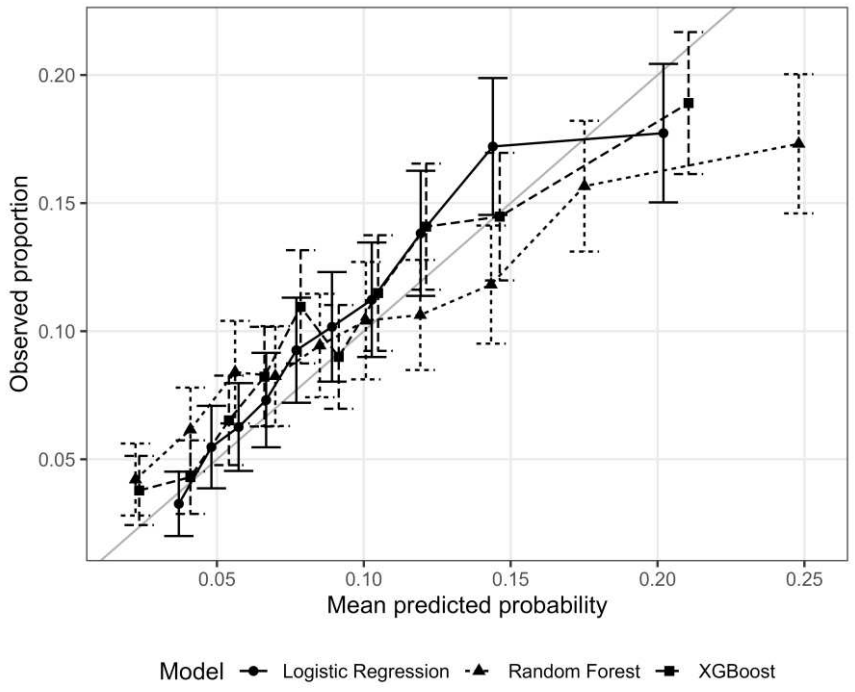


Figure 4A. Independent test-set calibration plot for logistic regression, random forest, and XGBoost, based on decile-binned predicted probabilities (error bars: 95% CI).

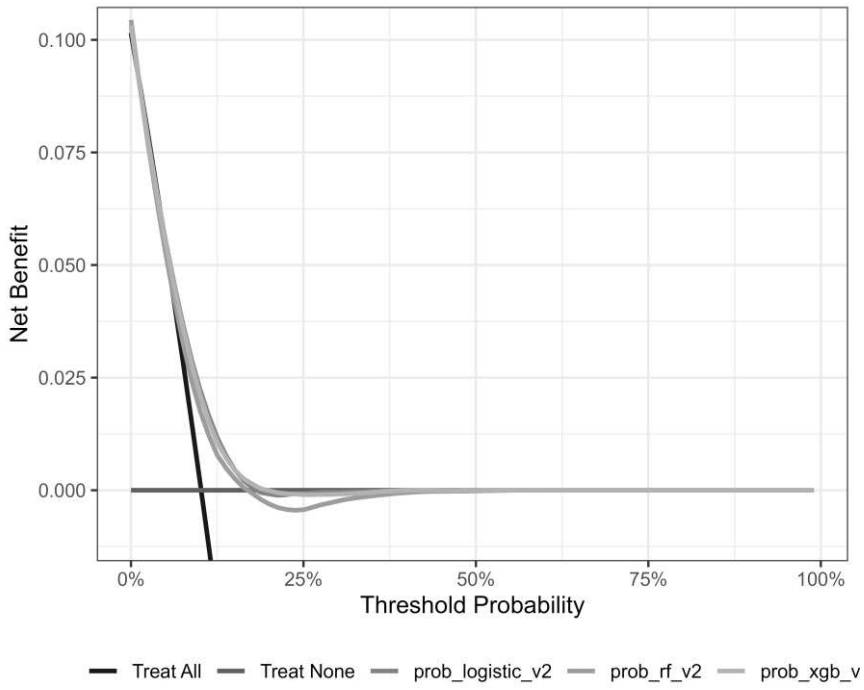


Figure 4B. Decision curve analysis comparing logistic regression, random forest, and XGBoost against treat-all and treat-none strategies.

SHAP explainability analysis

SHAP analysis of the XGBoost model identified age, waist circumference, and blood urea nitrogen as the top three predictors based on mean absolute SHAP values—indicating their greatest overall influence on individual risk predictions. These were followed by hypertension, dietary potassium intake, sex, and diabetes mellitus. Directionally, higher age, greater waist circumference, presence of hypertension or diabetes, male sex, and current smoking status were associated with positive SHAP contributions—i.e., increased predicted probability of kidney stone history—aligning consistently with the effect directions and magnitudes observed in the multivariable logistic regression model (Figure 5).

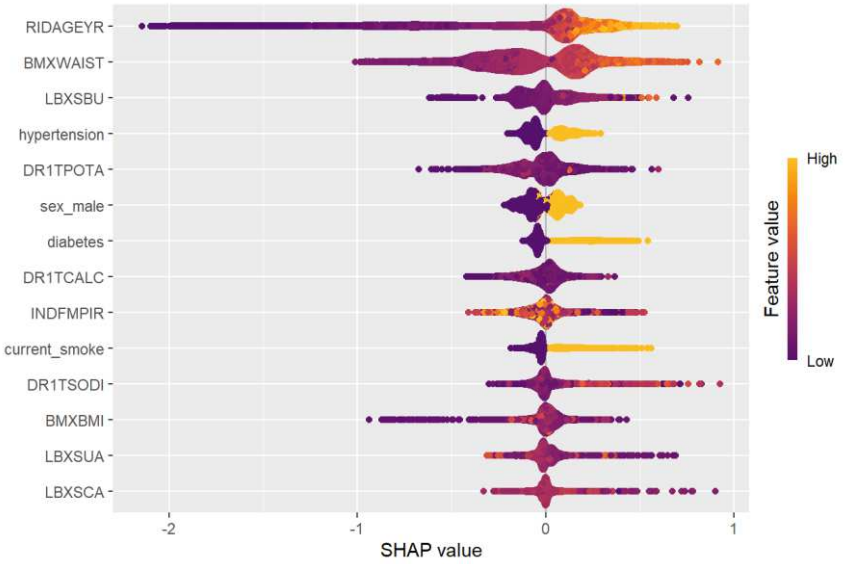


Figure 5. SHAP summary (beeswarm) plot for the XGBoost model. Each point represents one participant; color indicates the relative value of that variable (yellow = high, purple = low), and horizontal position indicates the SHAP value (impact on predicted probability).

Nomogram

A nomogram was constructed from the final logistic regression model to provide an interpretable, bedside tool for estimating an individual's probability of having a kidney stone history based on the 14 model variables (Supplementary Figure S1).

Survey-weighted sensitivity analysis

Odds ratios from the survey-weighted analysis were broadly consistent in direction and magnitude with the primary unweighted analysis (Table 4). The association for male sex was somewhat stronger in the weighted analysis (OR 1.357 vs. 1.223), while the association for blood urea nitrogen was attenuated and no longer statistically significant (weighted 95% CI 0.998–1.019).

Table 4. Comparison of unweighted and survey-weighted odds ratios for selected variables

Variable	Unweighted OR (95% CI)	Survey-weighted OR (95% CI)
Male sex	1.223 (1.078–1.387)	1.357 (1.151–1.602)
Waist circumference	1.025 (1.016–1.034)	1.018 (1.009–1.028)
Blood urea nitrogen	1.012 (1.003–1.021)	1.008 (0.998–1.019)
Hypertension	1.243 (1.105–1.397)	1.366 (1.173–1.590)
Diabetes	1.357 (1.182–1.554)	1.376 (1.163–1.628)
Current smoking	1.212 (1.061–1.382)	1.163 (1.018–1.328)

Discussion

In this nationally representative, pooled analysis of 25,560 US adults across six NHANES cycles (2007–2018), older age, male sex, greater waist circumference, elevated blood urea nitrogen, hypertension, diabetes mellitus, and current smoking were independently associated with self-reported lifetime kidney stone history. A modest inverse association with BMI—observed only in the multivariable logistic model—was likely attributable to multicollinearity with waist circumference, as confirmed by variance inflation factor assessment. Logistic regression demonstrated discrimination performance statistically equivalent to XGBoost and significantly superior to random forest, coupled with optimal calibration and the most robust clinical net benefit across the clinically relevant threshold probability range.[14-16]

A methodological note on smoking, confounding, and variable definition

An initial analysis relying solely on the NHANES questionnaire item for lifetime smoking of ≥100 cigarettes (SMQ020)—a variable frequently but imprecisely labeled "smoking status" in secondary analyses—yielded an implausibly high prevalence estimate (>40%), markedly exceeding contemporary national estimates of current smoking prevalence among US adults.[18] To address this misclassification, we constructed a refined current-smoking variable by integrating SMQ020 with current smoking frequency (SMQ040), thereby distinguishing never, former, and current smokers per standard epidemiologic practice. Using this validated definition, the unadjusted prevalence of current smoking was nearly identical between participants with and without kidney stone history (20.4% vs. 20.4%, P = 1.000). Nevertheless, current smoking remained independently and significantly associated with stone history after full multivariable adjustment (OR 1.212; 95% CI 1.078–1.364), indicating that the crude null association was obscured by confounding—most notably by age, as the stone-history group was significantly older. This pattern underscores two critical methodological imperatives: first, the necessity of precise, behaviorally grounded variable definitions to avoid outcome misclassification; and second, the essential role of multivariable adjustment in uncovering true associations masked by confounding in cross-sectional survey data.

Comparison with prior literature

The association between advancing age, male sex, and greater adiposity (as reflected by waist circumference) with kidney stone history is consistent with prior epidemiological work, including a large cohort analysis reporting that waist circumference partly mediates the excess kidney stone risk observed in men [6]. Similarly, the independent association of hypertension with stone history observed in our cohort aligns with a recent

NHANES-based analysis (n=21,740) reporting an adjusted OR of 1.36 (95% CI 1.19–1.56) for the association between hypertension and kidney stone prevalence [7], while the association with diabetes is consistent with another NHANES 2007–2018 analysis (n=21,177) demonstrating that diabetes partially mediates the relationship between central obesity and kidney stone prevalence [8]. Similar NHANES-based cross-sectional analyses have linked overall cardiovascular health metrics to kidney stone prevalence [30]. Prior NHANES-based work has also reported an association between active cigarette smoking and kidney stones using a properly defined current-smoking variable [9], consistent with our corrected analysis.

BMI exhibited a modest inverse association with kidney stone history in the fully adjusted multivariable model—after accounting for waist circumference—while both anthropometric measures displayed elevated variance inflation factors (VIF = 7.69 for BMI; VIF = 7.81 for waist circumference), confirming substantial multicollinearity. This finding aligns with established evidence demonstrating that, when jointly modeled, waist circumference—but not BMI—retains independent explanatory power for obesity-related cardiometabolic risk [10,19] Consequently, the directionally counterintuitive BMI coefficient in our model is best interpreted as a statistical artifact of collinearity rather than biologically meaningful evidence of a protective effect; its magnitude and sign should therefore not be overinterpreted in isolation..

Serum uric acid did not reach statistical significance in the adjusted model (OR 0.964, 95% CI 0.925–1.005), despite its established biological role in uric acid stone formation. This may reflect the fact that our outcome variable captured all self-reported kidney stones without distinguishing stone composition (e.g., calcium oxalate versus uric acid stones), diluting any composition-specific association.

Comparison of logistic regression and machine learning models

A key contribution of this study is the rigorous, multidimensional comparison of logistic regression against two ensemble machine learning algorithms—random forest and XGBoost—evaluated across three complementary performance dimensions: discrimination (assessed by AUC with formal pairwise DeLong's test), independent test-set calibration (intercept and slope of logistic calibration regression), and clinical utility (quantified via decision curve analysis across a clinically relevant 5–35% threshold range). Logistic regression and XGBoost demonstrated statistically equivalent discrimination (DeLong P = 0.591), whereas random forest was significantly inferior on all three metrics (all $P < 0.001$). This pattern aligns with findings from a recent systematic review of 71 studies evaluating machine learning versus logistic regression for binary clinical prediction tasks using structured, tabular data, which reported no consistent performance advantage for machine

learning methods—and, in many cases, poorer calibration and reduced generalizability [11]. Given its statistically comparable discrimination, superior calibration (intercept −0.147, slope 0.892), and full model transparency—operationalized through a validated nomogram—logistic regression represents the most pragmatic and clinically deployable approach for characterizing prevalent kidney stone history in routine practice or population-level screening. In contrast, random forest exhibited systematic overconfidence, a limitation that, if unaddressed, could lead to clinically misleading risk stratification and suboptimal decision-making [28, 29, 31].

Clinical and public-health implications

The nomogram developed in this study provides a practical, fully interpretable clinical tool for estimating the probability that an individual has experienced at least one kidney stone during their lifetime—based on routinely available demographic, anthropometric, and laboratory variables. Critically, because the outcome reflects a self-reported, prevalent condition (i.e., lifetime history) rather than an incident event, the nomogram is appropriately interpreted as a tool for prevalence characterization, not prospective risk prediction [22]. Its intended use is to support clinical decision-making in real-world settings—for example, by identifying individuals with elevated likelihood of prior stone disease to guide targeted metabolic evaluation, imaging, or preventive counseling [23]. It does not estimate the probability of future stone formation; such interpretation would require prospective longitudinal validation in cohorts with incident stone outcomes and is beyond the scope of this cross-sectional analysis [24].

Strengths and limitations

This study has several methodological strengths. First, it leverages a large, nationally representative pooled sample comprising 25,560 adults across six NHANES cycles (2007–2018), enhancing statistical power and generalizability. Second, variable selection was conducted systematically using LASSO logistic regression with 10-fold cross-validation, and a prespecified sensitivity analysis based on the lambda.1se criterion ensured robustness to model complexity. Third, key variables—including current smoking status and serum glucose—were explicitly redefined and validated against NHANES protocol documentation to minimize misclassification. Fourth, multicollinearity among predictors was formally assessed using variance inflation factors (VIF), with all retained variables exhibiting VIF < 10. Fifth, model evaluation followed a comprehensive, multimetric framework: discrimination was quantified by AUC with pairwise DeLong's tests; calibration was assessed in the

independent test set via logistic calibration regression (intercept and slope); and clinical utility was evaluated using decision curve analysis across the 5–35% threshold range. Sixth, SHAP-based explainability analysis provided transparent, instance-level interpretation of the XGBoost model. Finally, a survey-weighted sensitivity analysis—incorporating six-cycle examination weights, pseudo-strata, and pseudo-PSUs—confirmed the consistency of primary findings under complex-survey assumptions.

Several limitations warrant careful consideration. First, the cross-sectional design precludes causal inference. Moreover, because several predictors—including blood urea nitrogen, hypertension, and diabetes—were measured concurrently with the self-reported lifetime outcome, reverse causation remains plausible: for example, elevated blood urea nitrogen may reflect subclinical renal impairment secondary to prior stone disease rather than a true antecedent risk factor. Second, kidney stone history relied on a single self-reported item ("Ever had kidney stones?"), introducing potential recall bias and limiting granularity regarding stone composition, episode frequency, or recurrence status. Third, dietary exposures were estimated from a single 24-hour dietary recall, which—while standardized and rigorously administered—may not fully capture long-term habitual intake, particularly for episodically consumed foods or nutrients. Fourth, although the survey-weighted sensitivity analysis yielded broadly concordant estimates, the primary analysis used unweighted data; thus, prevalence and association measures reflect the pooled analytic sample and should not be interpreted as definitive nationally representative parameters without appropriate weighting. Fifth, internal validation employed a single train/test split within the same NHANES cohort; external validation in an independent, geographically or temporally distinct population would be necessary to confirm transportability and real-world generalizability. Sixth, the modest discrimination observed across all models (AUC 0.622–0.641) suggests that clinically relevant determinants—including urinary calcium/oxalate excretion, citrate levels, genetic variants, medication use, or detailed dietary patterns—remain unaccounted for in this analysis.

Conclusion

Among US adults, older age, male sex, greater waist circumference, elevated blood urea nitrogen, hypertension, diabetes, and current smoking were factors independently associated with a history of kidney stones. Logistic regression demonstrated discrimination comparable to XGBoost, significantly superior discrimination compared with random forest, the best calibration among the three models, and the most stable clinical net benefit, while remaining fully interpretable through an accompanying nomogram [33]. Given the cross-sectional design and lifetime, self-reported outcome, these findings characterize correlates of prevalent kidney stone history rather

than predictors of future risk, and should be externally validated before clinical application.

Declarations

Ethics approval and consent to participate: This study used publicly available, de-identified data from the National Health and Nutrition Examination Survey (NHANES). NHANES protocols were approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided written informed consent at the time of data collection. No additional ethical approval was required for this secondary analysis of de-identified, publicly available data.

Consent for publication: Not applicable. This study used de-identified, publicly available NHANES data and contains no individual participant details, images, or videos.

Competing interests: The authors declare that they have no competing interests.

Funding: This research received no external funding.

Authors' contributions: CL: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft. HW: Data curation, Visualization. DZ: Data curation, Validation. YW: Conceptualization, Methodology, Supervision, Writing – review & editing. All authors have verified the results and read and approved the final manuscript and agreed to submit it.

Acknowledgements: Not applicable.

Data availability: The datasets supporting the conclusions of this article are available in the National Health and Nutrition Examination Survey (NHANES) repository, https://wwwn.cdc.gov/nchs/nhanes/. Analysis code is available from the corresponding author upon reasonable request.

Generative AI statement: During the preparation of this manuscript, the authors used AI-assisted tools to improve the readability and language of the text. All content generated with the assistance of AI tools was reviewed and edited by the authors, who take full responsibility for the integrity and accuracy of the published work.

Clinical trial number: not applicable.

References

[1] Scales CD Jr, Smith AC, Hanley JM, Saigal CS, . Prevalence of kidney stones in the United States. Eur Urol. 2012;62(1):160-5.

[2] Abufaraj M, Xu T, Cao C, et al. Prevalence and Trends in Kidney Stone Among Adults in the USA:

Analyses of National Health and Nutrition Examination Survey 2007-2018 Data. Eur Urol Focus. 2021;7(6):1468-1475.
[3] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.

[4] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
[5] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74.

[6] Ferraro PM, Taylor EN, Curhan GC. Factors associated with sex differences in the risk of kidney stones. Nephrol Dial Transplant. 2023;38(1):177-183.
[7] He S, Chen YW, Hu LY, Chen QK, Wang Y, Liu SY. Association between hypertension and kidney stones: a cross-sectional analysis from NHANES and Mendelian randomization. Aging Male. 2025;28(1):2571937.
[8] Su Y, Wu R, Yuan H, He J, Pan H. A study on the relationship between the weight-adjusted waist index and kidney stone prevalence mediated by diabetes: data from NHANES 2007-2018. Sci Rep. 2025;15(1):19116.
[9] Huang Y, Wang H, Xu C, Zhou F, Su H, Zhang Y. Associations between smoke exposure and kidney stones: results from the NHANES (2007-2018) and Mendelian randomization analysis. Front Med (Lausanne). 10:1218051.
[10] Janssen I, Katzmarzyk PT, Ross R. Waist circumference and not body mass index explains obesity-related health risk. Am J Clin Nutr. 2004;79(3):379-84.
[11] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 110:12-22.
[12] Galán-Llopis JA, Sánchez-Pellicer P, Navarro-López V. Role of microbiome in kidney stone disease. Curr Opin Urol. 2023;33(2):84-89.

[13] Zhang M, Zhang J, Cui Y, Xing Z. Inverted V-shaped association between atherogenic index of plasma and kidney stone risk: results from NHANES 2011-2018. Transl Androl Urol. 2025;14(4):953-963.

[14] Stamatelou K, Goldfarb DS. Epidemiology of Kidney Stones. Healthcare (Basel). 2023;11(3).

[15] Stritt K, Plouvin M, Younes SE, et al. Epidemiology and risk factors of urolithiasis: insights from SKIPOGH, a population-based cohort study in Switzerland. Clin Kidney J. 2026;19(2):sfag002.

[16] Alshubaili AM, Alotaibi AF, Alsaleh KA, et al. The Prevalence of Nephrolithiasis and Associated Risk Factors Among the Population of the Riyadh Province, Saudi Arabia. Cureus. 2024;16(3):e55870.

[17] Hill AJ, Basourakos SP, Lewicki P, et al. Incidence of Kidney Stones in the United States: The Continuous National Health and Nutrition Examination Survey. J Urol. 2022;207(4):851-856.

[18] Crivelli JJ, Maalouf NM, Paiste HJ, et al. Disparities in Kidney Stone Disease: A Scoping Review. J Urol. 2021;206(3):517-525.

[19] Garg M, Joseph P, Johnson H, et al. Obesity and BMI score as risk factors for urolithiasis: a systematic review over 30 years. Curr Opin Urol. 2025;35(4):338-346.

[20] Song R, Wu K, Ma M, et al. Association between oxidative balance score and kidney stones: data from the national health and nutrition examination survey (NHANES). BMC Nephrol. 2024;25(1):190.

[21] Shen Y, Zhu Z, Bi X, et al. Association between insulin resistance indices and kidney stones: results from the 2015-2018 National Health and Nutrition Examination Survey. Front Nutr. 2024;11:1444049. doi:10.3389/fnut.2024.1444049.

[22] Park DJ, Kim BS, Kwon SO, et al. Clinical characteristics of surgically managed patients with asymptomatic renal stones: Comparison of patients with symptomatic renal stones. Investig Clin Urol. 2023;64(2):161-167.

[23] Wang X, Xu N, Xu M. Pertinence between Preoperative Renal Function and Composition of Renal Calculi in Elderly Patients: A Retrospective Analysis. Arch Esp Urol. 2024;77(10):1125-1132.

[24] Zhou X, Jin K, Qiu S, et al. Associations of exposure to polycyclic aromatic hydrocarbons and kidney stones in U.S. general population: results from the National Health and Nutrition Examination Survey 2007-2016. World J Urol. 2022;40(2):545-552.

[25] Zhou X, Jin K, Qiu S, et al. Associations of exposure to polycyclic aromatic hydrocarbons and kidney stones in U.S. general population: results from the National Health and Nutrition Examination Survey 2007-2016. World J Urol. 2022;40(2):545-552.

[26] Kayar K, Kayar R, Tuncel KG, et al. Development of a calibrated logistic regression clinical risk score for

predicting postoperative sepsis following retrograde intrarenal surgery. World J Urol. 2026;44(1).

[27] Wei F, Sun M, Tang T, Yang F, Dai X, Zhou Y. Establishment of a nomogram model and scoring system for kidney stone prediction based on diet and lifestyle. Medicine (Baltimore). 2025;104(43):e45351.

[28] Qiu P, Liu J, Wan F, et al. A predictive model for postthrombotic syndrome in proximal deep vein thrombosis patients. Ann Transl Med. 2021;9(7):558.

[29] Oearsakul T, Tunthanathip T. Development of a nomogram to predict the outcome of moderate or severe pediatric traumatic brain injury. Turk J Emerg Med. 2022 Jan-Mar;22(1):15-22.

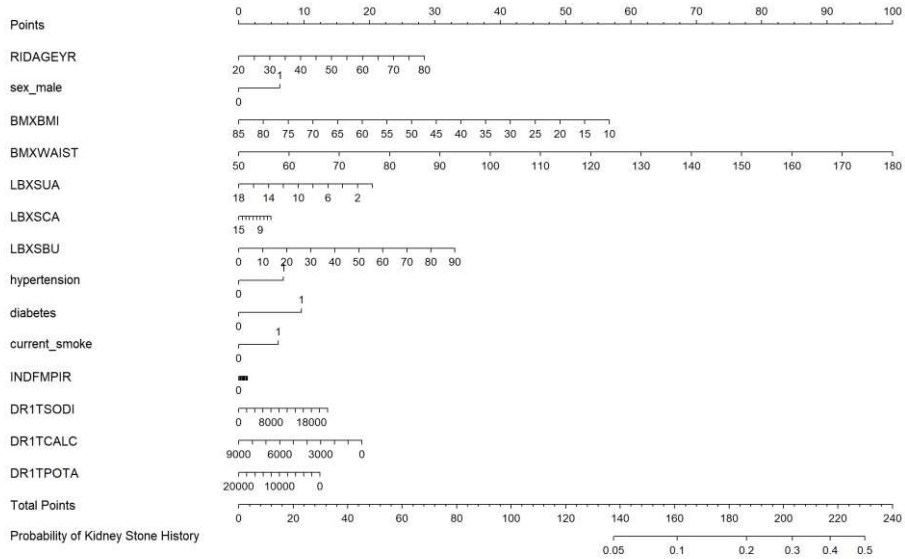
[30] Yin S, Yang Z, Zhu P, et al. Association between Two Cardiovascular Health Algorithms and Kidney Stones: A Nationwide Cross-sectional Study. Eur Urol Open Sci. 73:8-16.

[31] Fan D, Liu H, Han Y, Gao Z, Gu Y. Machine learning algorithms for predicting stone residue and recurrence after lateral decubitus percutaneous nephrolithotomy. Medicine (Baltimore). 2025;104(38):e44750.

[32] Fan D, Liu H, Han Y, Gao Z, Gu Y. Machine learning algorithms for predicting stone residue and recurrence after lateral decubitus percutaneous nephrolithotomy. Medicine (Baltimore). 2025;104(38):e44750.

[33] Sánchez C, Larenas F, Arroyave JS, et al. Artificial Intelligence in Urology: Application of a Machine Learning Model to Predict the Risk of Urolithiasis in a General Population. J Endourol. 2024;38(8):712-718.

Supplementary Figure



Supplementary Figure S1. Nomogram for estimating the probability of kidney stone history, based on the final logistic regression model (14 variables).