Original Paper

Machine Learning Models Combining Routine Urine Dipstick Results With Age, Sex, and BMI to Triage Serum Urate Testing: Development and External Validation Study

Eun Chan Jang[1,2], MPH; Pado Song[2,3], MD; Joung Eun Kim[2,3], MD; Myung Hwa Han[1,2], MPH; Hwigyeong Han[2], MPH; Kiheon Lee[1,2,3,4], MD, PhD

[1]Department of Health Science and Technology, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea

[2]Department of Family Medicine, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

[3]Health Promotion Center, Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, Republic of Korea

[4]Department of Family Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Kiheon Lee, MD, PhD
Department of Family Medicine
Seoul National University Bundang Hospital
82, Gumi-ro 173 Beon-gil, Bundang-gu
Seongnam-si, Gyeonggi-do, 13620
Republic of Korea
Phone: 82 31 787 7801
Email: keyhoney@gmail.com

Abstract

Background: Serum urate testing requires venipuncture; urine dipstick testing is inexpensive and routine. Whether dipstick results add information on hyperuricemia beyond age, sex, and BMI is unknown.

Objective: This study aims to develop and externally validate machine learning models estimating the probability of hyperuricemia from 9 dipstick fields, age, sex, and BMI, and to quantify what the dipstick adds.

Methods: The development dataset included 129,535 adults with first urine tests and same-day serum urate and BMI measurements at a Korean tertiary hospital (2010-2025); 21,124 had urate levels ≥7.0 mg/dL, and 81.6% (n=105,704) had a same-day health checkup. Calibrated 12-input ensembles of Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost) were trained on a person-level 80% split (103,628 persons) for serum urate ≥7.0 mg/dL (primary) and ≥9.0 mg/dL (secondary). They were tested in the remaining 25,907 persons (4225 events at ≥7.0 mg/dL) and in 49,987 adults in the KNHANES (Korea National Health and Nutrition Examination Survey; 2016-2024; 5068 events). Inputs, KNHANES coding, split, and training procedure were fixed before fitting. We assessed discrimination, calibration, operating points by sex, decision curves, and increments over age-sex-BMI models.

Results: At ≥7.0 mg/dL, the area under the receiver operating characteristic curve (AUROC) was 0.825 (95% CI 0.820-0.831) internally and 0.845 (95% CI 0.840-0.849) externally, but 0.710 in men and 0.795 in women. Across 4 model classes, the dipstick added 0.017 to 0.024 to the external AUROC of age-sex-BMI models (0.814-0.823), or 3.1 cases per 1000 screened with top-10% referral. Externally, the models overpredicted (observed-to-expected ratio 0.73, 95% CI 0.71-0.75); an intercept update fitted in half of KNHANES restored it to 1.00 in the other half. The internally derived 80%-sensitivity threshold flagged 29.8% (14,916/49,987) of adults, 97.4% (14,533/14,916) of them men; sensitivity was 77.1% (3907/5068) overall but 10.1% (53/527) in women, and positive predictive value (PPV) was 26.2% (3907/14,916). Flagging all men gave 89.6% (4541/5068) sensitivity and slightly less net benefit than the model at threshold probabilities up to 10%. At ≥9.0 mg/dL, external results were AUROC 0.860, increment 0.052 to 0.074 (1.6 cases per 1000), observed-to-expected ratio 0.48, scaled Brier score 0.002 (no better than predicting the prevalence), and PPV 2.6% at 82% sensitivity.

Conclusions: Routine dipstick results add a small increment to age, sex, and BMI for identifying serum urate ≥7.0 mg/dL. Ranking transferred from the hospital to a national survey sample, but absolute risk did not, and at the tested thresholds the models mainly selected men. Both cohorts comprised people who had blood drawn, so triage of serum urate testing in urine-only encounters,

XSL•FO
RenderX

in which BMI is usually missing, remains an untested hypothesis. Before any use, the models would need local recalibration and evaluation in the intended setting.

(JMIR Med Inform 2026;14:e107515) doi: 10.2196/107515

KEYWORDS

calibration; clinical prediction model; decision curve analysis; external validation; gout; hyperuricemia; KNHANES; Korea National Health and Nutrition Examination Survey; machine learning; triage; urine dipstick

Introduction

Hyperuricemia affects an estimated 20% to 25% of adult men and 5% to 10% of adult women, with prevalence rising alongside obesity and dietary change [1,2]. Beyond its role in gout [3], it is associated with cardiovascular disease [4], chronic kidney disease [5], nonalcoholic fatty liver disease [6], and mortality [7]. Urate-lowering therapy is not recommended for asymptomatic hyperuricemia [8], but detection can prompt lifestyle counseling and attention to cardiovascular and renal risk. National survey data show that hyperuricemia remains common in adults [9].

Serum urate is measured in blood, and ≥7.0 mg/dL is the usual cutoff, close to the monosodium urate solubility limit of about 6.8 mg/dL [8,10]. When blood is drawn, adding serum urate to the sample is cheap and definitive. Many encounters, however, collect urine without blood, for example, for urinary symptoms, and uric acid status is then not assessed.

Urine dipstick testing is noninvasive, inexpensive, and routine. Several mechanisms could link dipstick results to serum urate: urinary pH determines urate solubility and excretion, proteinuria reflects glomerular damage that may impair urate clearance, and specific gravity reflects hydration [11,12]. Age, sex, and BMI are strong determinants of serum urate, partly through insulin-mediated reduction in renal urate excretion [13,14].

We asked whether routine dipstick results, combined with age, sex, and BMI, carry information on serum urate that could help decide who should have a confirmatory blood test. Machine learning can capture nonlinear relationships among such inputs [15], but the practical value of a model depends on what the dipstick adds to demographic information, on calibration in the target population, and on the number of tests needed per case found. We developed models for serum urate ≥7.0 mg/dL (primary) and ≥9.0 mg/dL (secondary; a level at which gout incidence rises steeply [16,17]) and report their external validation, their increment over age, sex, and BMI, and their operating characteristics overall and by sex. A ≥10.0 mg/dL model is reported as exploratory.

Methods

Ethical Considerations

The institutional review board (IRB) of Seoul National University Bundang Hospital approved the study (IRB number B-2509-994-106) and waived informed consent for the retrospective analysis of deidentified records. KNHANES is conducted by the Korea Disease Control and Prevention Agency (KDCA) with participants' informed consent. The 2016 and 2017 surveys were conducted without IRB review, following the KDCA IRB's opinion that KNHANES is research carried out directly by the state for the public good under the Bioethics and Safety Act; the 2018-2024 surveys were approved by the KDCA IRB (2018-01-03-P-A, 2018-01-03-C-A, 2018-01-03-2C-A, 2018-01-03-5C-A, 2018-01-03-4C-A, 2022-11-16-R-A, and 2022-11-16-R-03) [18-20]. KNHANES data are publicly available and deidentified, and their use required no separate approval. The study followed the Declaration of Helsinki.

Study Design

This retrospective study used hospital records to develop 3 prediction models, 1 for each serum urate threshold, tested them in held-out persons from the same hospital, and validated them externally in a national health survey, the KNHANES (Korea National Health and Nutrition Examination Survey). It is reported according to the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis extension for Artificial Intelligence) statement [21] (TRIPOD+AI checklist provided in Multimedia Appendix 1) and was informed by the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [22] and RECORD (Reporting of Studies Conducted Using Observational Routinely Collected Health Data) [23] statements. No protocol was prepared or registered, and patients and the public were not involved (Tables S1-S22 and Figures S1-S4 in Multimedia Appendix 2).

Data Sources and Participants

Development Cohort

The data came from the laboratory information system of a tertiary university hospital in the Seoul metropolitan area. For each person, we used the first urine dipstick test recorded between January 2010 and August 2025 (620,506 persons). The target population of the models is adults aged 19 years or older, as in the external validation cohort; 572,330 persons were aged 19 years or older at the urine test, and 48,173 younger persons and 3 with an implausible age were excluded. Age was calculated in whole years (days since birth divided by 365.25, rounded down); recomputed as exact completed years, no person in the final cohort was aged younger than 19 years. The urine test was linked by patient identifier to blood tests within 14 days before or after it, and 534,887 adults had at least 1 such blood test.

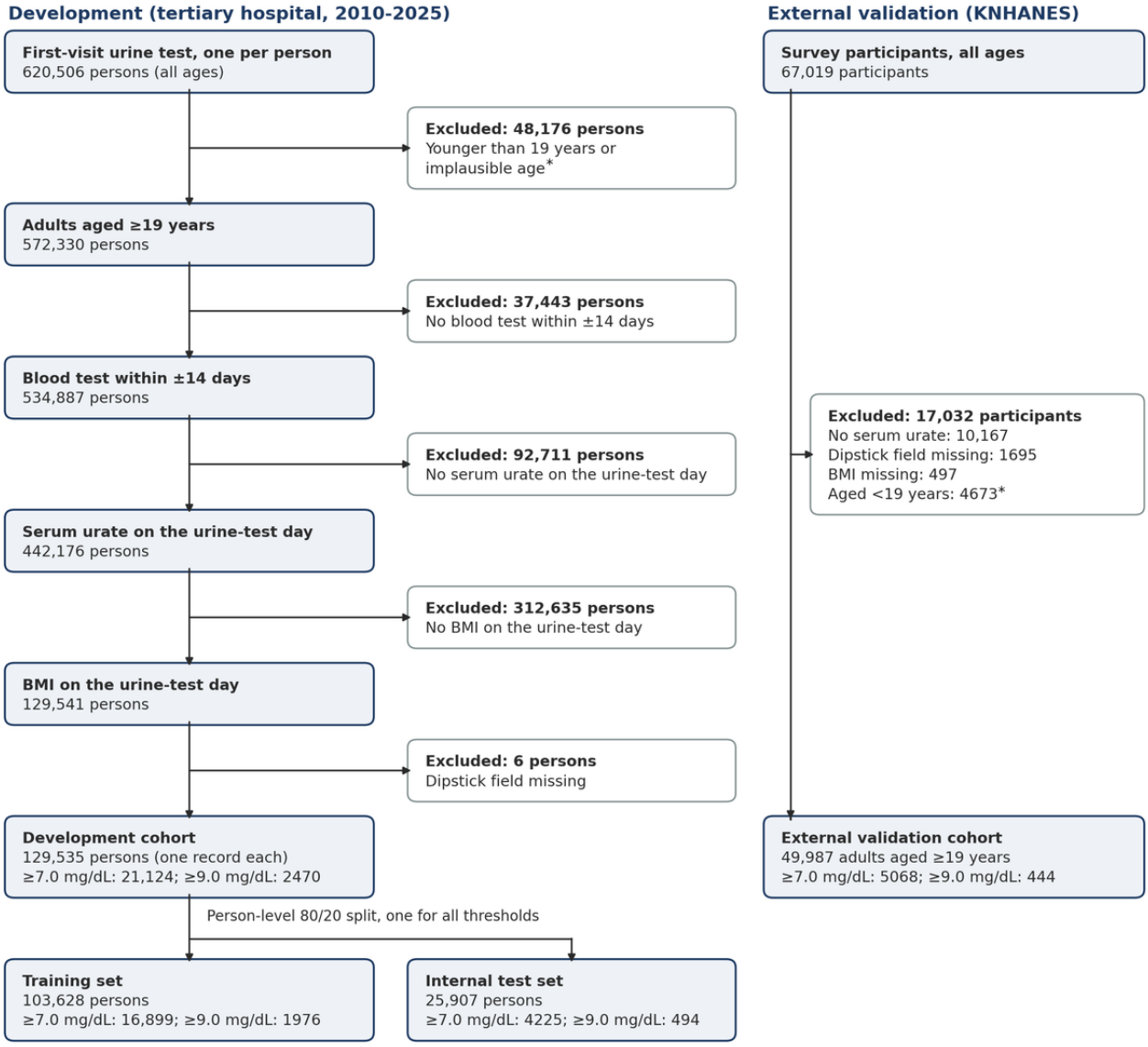
Serum urate and BMI had to be recorded on the day of the urine test, so that the outcome and every input describe the same encounter; values from other days were not used (in these data, every health-checkup BMI from another day came from a later

XSL•FO
RenderX

checkup and would not have been available at the urine test). Serum urate was measured on that day in 442,176 adults. When it was measured more than once that day, the earliest result was used, and results reported at the same time were averaged; results reported as "<1.0 mg/dL" were counted as 0.5 mg/dL (40 persons in the development cohort). BMI was recorded on that day for 129,541 of these adults. It was taken from a health checkup held on that day (calculated from height and weight measured at the checkup) or, if there was none, calculated from height and weight in the nursing record of that day; nursing-record height outside 50 cm to 220 cm or weight outside 10 kg to 300 kg, and BMI outside 10 kg/m^2 to 70 kg/m^2, were treated as missing. Finally, all 9 dipstick fields used as inputs and sex had to be recorded, which gave a development cohort of 129,535 adults (Figure 1 and panel A in Table S1 in Multimedia Appendix 2). BMI came from the health checkup for 105,704 (81.6%) persons and from the nursing record for 23,831 (18.4%).

The unit of analysis is the person: each person contributes 1 record, the first urine test with the serum urate and BMI of that day. The data do not contain information on encounter type; the only marker of setting is a health-checkup record dated on the day of the urine test. Because each person's first urine test was used, the early years include patients with established hyperuricemia, and persons whose first test lacked a same-day serum urate or BMI were excluded even if a later test had them.

Figure 1. Participant flow for the development cohort, its person-level split into training and internal test sets, and the KNHANES (Korea National Health and Nutrition Examination Survey) external validation cohort. Persons aged younger than 19 years, outside the adult target population of the models, were excluded from both cohorts. Serum urate and BMI are those recorded on the day of the urine test. *Outside the adult target population.



Training and Internal Test Sets

The development cohort was split once at random, at the person level, into a training set (103,628/129,535, 80%) and an internal test set (25,907/129,535, 20%), stratified on serum urate bands (<7.0, 7.0 to <9.0, 9.0 to <10.0, and ≥10.0 mg/dL; seed 42). The 3 thresholds share this split, so the internal test set holds the same persons for every model, and no person is in both sets. The training set held 16,899 events at ≥7.0 mg/dL and 1976 at ≥9.0 mg/dL, and the internal test set held 4225 and 494 (16.3% and 1.9% in both sets). All comparison models were fitted on the same training set as the models, except in the temporal

XSL•FO
RenderX

holdout analysis (Sensitivity Analyses for Coding, Linkage, and Time subsection).

External Validation Cohort

KNHANES is an annual cross-sectional survey of the Korean population [18]; serum urate has been part of its laboratory tests since 2016 [19]. The analysis file concatenated the nine annual files for 2016-2024 (67,019 participants of all ages). We required serum urate (56,852), the 9 dipstick fields used as inputs (55,157), measured BMI (54,660), and age of 19 years or older, which gave 49,987 adults (Figure 1). KNHANES age is top-coded at 80 years. The main analysis is unweighted; survey weights were used in a sensitivity analysis, and we describe the sample as a national survey sample rather than a nationally representative one. No sample size calculation was performed, and all eligible persons were included in both cohorts; the external cohort's 5068 and 444 events at ≥7.0 and ≥9.0 mg/dL exceed the minimum of 100 events, preferably 200, recommended for external validation [24], although formal criteria [25] were not applied.

Input Features

The models use 12 inputs: 9 dipstick fields (glucose, protein, ketone, bilirubin, occult blood, urobilinogen, nitrite, pH, and specific gravity), age, sex, and BMI. The development laboratory also reports leukocyte esterase, but it was not used as an input because KNHANES does not measure it. In the development data, semiquantitative fields were coded 0 (negative), 0.5 (trace), 1 (1+), 2 (2+), and 3 (3+ or more); the development laboratory reports no trace level for bilirubin and codes a normal urobilinogen as 0.5. Nitrite was binary (negative or positive), and pH and specific gravity were continuous. Specific gravity was recorded in steps of 0.005 up to 1.030 in the development data, consistent with a reagent-pad reading, but in steps of 0.001 up to 1.050 in KNHANES, where it is measured by reflection refractometry and the other dipstick fields are read by reflectance from Sysmex MEDITAPE UC-9A strips on a UC-3500 analyzer (2022-2024) [20]; the development laboratory's instrument is not documented (Tables S2 and S3 in Multimedia Appendix 2).

KNHANES records semiquantitative results as codes 0 to 5 (negative, trace, 1+, 2+, 3+, and 4+). Before the models were fitted, the KNHANES coding was fixed to a label-matched mapping (X1): each code was assigned to the development grade with the same label (for example, bilirubin "+" to 1+, coded 1), and 4+ was merged into 3 because the development laboratory's top category is "3+ or more." A normal or trace urobilinogen was coded 0.5, the development code for a normal result; any positive nitrite code was coded 1; and pH and specific gravity were used as recorded. The tree-based learners of the models use only the order of the values of each input, so the unequal numeric intervals of the codes do not affect them; what matters is that each KNHANES grade is assigned to the development grade with the same label. The intervals do affect the logistic comparison models, whose dependence on the coding is reported. Equivalence of grades across strip manufacturers and analyzers is assumed, and no crosswalk or paired-strip study validates any mapping; alternative codings were therefore evaluated as sensitivity analyses (Tables S2 and S4 in Multimedia Appendix 2).

Outcomes

Three binary outcomes were defined from serum urate measured on the day of the urine test: ≥7.0 mg/dL (primary), aligned with the institutional reporting standard and the urate solubility limit [8,10]; ≥9.0 mg/dL (secondary), above which incident gout increases markedly [16,17]; and ≥10.0 mg/dL (exploratory), the institution's threshold for starting urate-lowering therapy, reported only in Tables S21 and S22 and Figure S4 in Multimedia Appendix 2 (Section S12). In KNHANES 2022-2024, serum urate was measured by an enzymatic colorimetric assay [20]; the assay of the hospital laboratory was not documented in the data available to us.

Model Development

For each threshold, the model is an equal-weight soft-voting ensemble of Light Gradient Boosting Machine (LightGBM) [26], Extreme Gradient Boosting (XGBoost) [27], and Categorical Boosting (CatBoost) [28], calibrated by isotonic regression [29,30], and it outputs the probability that serum urate is at or above the threshold. The 3 models were developed with the same procedure on the training set only. The hyperparameters of each learner were tuned by tree-structured Parzen estimator optimization (Optuna; sampler seed 42; 25 trials per learner), maximizing the mean area under the receiver operating characteristic curve (AUROC) of 3-fold stratified cross-validation with shuffling (seed 42). For every learner, the search covered 100 to 300 trees (in steps of 50), a maximum depth of 4 to 8, and a learning rate of 0.02 to 0.10 (log scale), together with learner-specific parameters: for LightGBM, the number of leaves (15-63), the minimum number of samples in a leaf (10-100), and L1 and L2 regularization (0.001-1.0, log scale); for XGBoost, row and column subsampling (0.7-1.0) and the minimum child weight (1-10); and for CatBoost, L2 leaf regularization (0.01-10, log scale; tuned values provided in Table S8 in Multimedia Appendix 2). Learner seeds were fixed at 42, and no class weighting or resampling was used. The tuned learners were combined by soft voting, and the ensemble was calibrated with 5-fold cross-fitting on the training set (stratified folds without shuffling): in each fold, the 3 learners were fitted on four-fifths of the training persons and the isotonic map on the remaining fifth, and the model averages the 5 calibrated ensembles (15 base learners in all). Each calibration fold contained 3379 to 3380 events at ≥7.0 mg/dL and 395 to 396 at ≥9.0 mg/dL. The cohort, split, inputs, KNHANES coding, and development procedure, including the search spaces, were fixed before the models were fitted or evaluated and recorded with checksums of the training data, the training and test identifiers, and the code. Predictions in the internal test set and in KNHANES were computed with the saved models after training was complete.

Statistical Analysis

Discrimination and Calibration

Discrimination was measured by the AUROC and the area under the precision-recall curve (AUPRC; average precision) [31], and calibration by the observed-to-expected (O:E) ratio, the

XSL•FO
RenderX

calibration intercept (with the slope fixed at 1) and slope, the Brier and scaled Brier scores (negative when less accurate than predicting the prevalence), and calibration plots in deciles of predicted risk with Wilson 95% CIs for the observed proportions [30,32]. Because isotonic calibration produces a step function, we counted predictions of exactly 0 and, as sensitivity analyses, floored isotonic outputs at 0.001 and compared isotonic calibration with 2 parametric calibrators fitted on the same fold learners and calibration folds: a Platt (sigmoid) calibrator and a logistic calibrator on the log-odds of the ensemble's predicted probability. External recalibration used an intercept update, logit(p') = a + logit(p), fitted on a random half of KNHANES and evaluated on the other half (100 splits, stratified on the outcome) [33]; logistic recalibration (intercept and slope) was fitted in the same way for comparison.

CIs and Comparisons

Unless stated otherwise, 95% CIs were calculated as percentile bootstrap intervals from 1000 resamples drawn with replacement at the full size of the evaluation set (seed 42). For example, the cases-per-1000 differences and the net-benefit comparisons in panel C in Table S13 in Multimedia Appendix 2 used 400 paired resamples. The models were held fixed, so CIs reflect sampling variability of the evaluation sets only. The resampling unit was the person (1 record per person in each cohort); resamples with a single outcome class were discarded, and subgroup CIs resampled within the subgroup. Paired AUROC differences used paired bootstrap resamples (primary) and the DeLong test [34]. Independent subsets (women and men, and persons with checkup and nursing-record BMI) were compared with the DeLong method for independent curves, and women and men also with a bootstrap in which each sex was resampled independently (2000 resamples). Analyses over 100 random half-splits report the mean and the 2.5th to 97.5th percentiles across splits.

Clinical Utility

Thresholds were chosen on the internal test set (Youden index, 80% and 90% sensitivity) and applied unchanged to KNHANES; the internal values are therefore apparent performance. As sensitivity analyses, thresholds were also chosen on out-of-fold training predictions (not fully out of sample, because each fold's isotonic map had been fitted on the persons it scored) and on a random half of KNHANES (applied to the other half; 100 random splits). We report sensitivity, positive predictive value (PPV), and the proportion flagged, overall and within each sex, also for sex-specific thresholds (80% sensitivity within each sex) and for referral of a fixed fraction of persons with the highest predicted risk (for example, the top 10%), and compare 2 reference strategies: testing all men, and an age-sex-BMI spline logistic model fitted on the training set, with thresholds chosen in the same way. Decision curves [35] (threshold probabilities 2%-50% at ≥7.0 mg/dL and 0.2%-10% at ≥9.0 mg/dL) were computed for the models as validated and after the intercept update, cross-fitted in KNHANES (fitted on one half and applied to the other, both ways; means over 100 random splits, with 95% CIs from bootstrap resamples of one split), and for the reference strategies; operating points were also read at fixed risk thresholds on the recalibrated scale.

Added Value of the Dipstick and Comparison With Simpler Models

We fitted models on the training set for 5 input sets: age and sex (2F); age, sex, and BMI (3F); the 9 dipstick fields (9F); the dipstick fields with age and sex (11F); and all 12 inputs (12F). Each set was fitted with 1 fixed LightGBM configuration, main-effects logistic regression, and, for the sets that contain age, logistic regression with cubic splines for age and BMI plus an age-by-sex term; LightGBM was also tuned per input set (12-configuration grid, by 3-fold cross-validated AUROC on the training set). All were evaluated in the internal test set and in KNHANES, and increments were also estimated within each sex in KNHANES. The clinical value of the dipstick was expressed as the additional cases detected per 1000 KNHANES adults screened when the adults with the highest predicted risk (top 1% to 20%) are referred, and as the difference in net benefit, comparing the fixed-configuration LightGBM models with all 12 inputs and with age, sex, and BMI. To test whether the ensemble is needed, it was compared with simpler 12-input models fitted on the same training set: the logistic models, LightGBM with the fixed and with the grid-tuned configuration, and the ensemble's own tuned LightGBM learner calibrated on the same isotonic folds, which isolates the contribution of XGBoost and CatBoost.

Missing Data

We report the missingness of the inputs and of same-day serum urate among the 534,887 adults whose first urine test had a blood test within 14 days, compared the characteristics of persons excluded only because BMI was not recorded on the urine-test day with those of included persons, and applied the Little missing completely at random test [36] to all these adults and to random subsamples of 1000 to 100,000 persons (10 of each size), interpreting only its rejection because the binary and ordinal inputs violate its normality assumption. The models were developed and evaluated in complete cases, because their intended use requires BMI measured at the encounter. Under a missing-at-random assumption, we evaluated the models in the 312,613 adults who had a same-day serum urate and complete dipstick fields, age, and sex but no BMI recorded on that day, with BMI multiply imputed (20 imputations from a Bayesian ridge regression with posterior draws; Rubin rules [37]) from age, sex, the dipstick fields, and serum urate, or, as a pessimistic bound, from age and sex only.

Verification Bias and Setting

Serum urate is not measured in every hospital encounter, so the development outcome is observed selectively [38,39]. Among the 534,831 adults whose first urine test had a blood test within 14 days and who had complete dipstick fields, age, and sex, we modeled the probability that serum urate was measured on the day of the urine test by logistic regression on the dipstick fields, 10-year age group, sex, and calendar year and, as a sensitivity analysis, among those with BMI recorded on that day. The internal AUROC and O:E ratio were reweighted by the inverse of this probability (weights truncated at the 99th percentile). We compared internal performance between persons with and without a same-day health checkup, who in this cohort are the persons with checkup and with nursing-record BMI, overall

XSL•FO
RenderX

and within age bands (19-64 years and ≥65 years), and described urine tests without a blood test and KNHANES participants without serum urate.

Subgroups, Nested Thresholds, and a Continuous-Urate Model

Performance was estimated by age group (19-39 years, 40-59 years, 60-79 years, and ≥80 years) and sex, in KNHANES (primary) and in the internal test set; the female-male difference in AUROC was also estimated among adults aged 40 to 69 years, whose age distributions are more similar in the 2 sexes. Because separately trained models can assign a higher probability to a stricter threshold, we counted such violations for each pair of thresholds in KNHANES and in the internal test set, which holds the same persons for the 3 models. As a sensitivity analysis, a LightGBM regression of log serum urate, fitted on the training set, derived P(urate ≥ t) for every threshold from 5-fold out-of-fold residuals, assuming constant variance (a variant that also modeled the variance is provided in panel C in Table S18 in Multimedia Appendix 2). To separate the continuous formulation from tuning and ensembling, it was compared with binary LightGBM models with the same fixed configuration, fitted on the same training set for each threshold.

Sensitivity Analyses for Coding, Linkage, and Time

KNHANES was scored under 4 alternative codings: M1, which codes bilirubin trace, "+", and "++" as 1+, 2+, and 3+ and the urobilinogen codes 0 to 5 as 0, 0.2, 0.5, 1, 2, and 3; M2, which differs from X1 only in coding a normal urobilinogen as 0; M3, which is M1 with a normal urobilinogen coded 0.5; and X2, which is X1 with specific gravity rounded to the nearest development grade (steps of 0.005, capped at 1.030). We also restricted KNHANES to specific gravity of 1.030 or below, set ketone results to negative (all results in KNHANES; trace or all results in the internal test set), and computed the population stability index of each input. For BMI, we compared checkup and nursing-record values recorded on the same day and examined terminal-digit preference in checkup and nursing-record height and weight. To assess performance over time, the ensemble was refitted with its tuned hyperparameters (not retuned; they had been tuned on a training set that included persons tested from 2022) on persons whose urine test was in 2010-2021 and evaluated in those tested in 2022-2025; fixed-configuration LightGBM (all 12 inputs; age, sex, and BMI) and main-effects logistic regression (all 12 inputs) were fitted in the same way and, as a reference, on the training set of the random split. Because the development laboratory changed its dipstick reporting around December 2021, with a transition from July 2021 to January 2022, this holdout mixes the passage of time with a change in measurement. The models were also evaluated by period of the urine test in the internal test set (2010-2021 and 2022-2025), by KNHANES survey year, and with KNHANES survey weights (CIs ignore the survey strata and primary sampling units).

Interpretation of the Ensemble

Input importance was estimated for the whole calibrated ensemble by permutation importance [40] (decrease in AUROC; 10 repeats) and by Shapley additive explanations (SHAP), computed with TreeSHAP [41] and averaged over all 15 base learners (all 49,987 KNHANES adults and all 25,907 internal test persons; per-learner ranks and probability-scale SHAP in a random sample of 1000 KNHANES adults), and effect shapes by calibrated partial dependence on specific gravity, pH, age, and BMI, overall and within each sex. To interpret the effect of specific gravity, we estimated its association with each outcome by logistic regression, crude and adjusted for sex, age, and BMI, in both cohorts.

Software

Analyses used Python 3.13.5 with *scikit-learn* 1.6.1, *LightGBM* 4.6.0, *XGBoost* 3.1.2, *CatBoost* 1.2.8, *Optuna* 4.8.0, *SHAP* 0.48.0, and *statsmodels* 0.14.4. Analysis code was drafted with a large language model agent (Claude Code; Anthropic) and reviewed by the authors (see the Generative AI Use section).

Results

Study Population and Setting

The development cohort comprised 129,535 adults (mean age 49.0, SD 13.5 years; men: n=67,443, 52.1%) and the external cohort comprised 49,987 KNHANES adults (mean age 52.5, SD 16.8 years; men: n=22,281, 44.6%; Figure 1 and Table 1). Serum urate ≥7.0 mg/dL was less common in KNHANES (5068/49,987, 10.1% vs 21,124/129,535, 16.3%), as was ≥9.0 mg/dL (444/49,987, 0.89% vs 2470/129,535, 1.9%).

In the development cohort, 105,704 (81.6%) persons had a health checkup on the day of the urine test, which also provided their BMI; the other 23,831, whose BMI came from the nursing record, were older (mean age 54.4, SD 16.9 vs 47.8, SD 12.2 years). Among the 534,887 adults whose first urine test had a blood test within 14 days, serum urate was measured on the urine-test day in 99.04% (105,957/106,981) of those with a same-day checkup and 78.6% (336,219/427,906) of the others, and BMI was recorded on that day in 99.47% (106,410/106,981) and 6.6% (28,434/427,906; Table S1 in Multimedia Appendix 2). Of the 572,330 adults with a first urine test, 37,443 (6.5%) had no blood test within 14 days; they were of similar age to those with a blood test (mean age 53.0, SD 17.5 vs 54.0, SD 16.1 years) but more often women (25,559/37,443, 68.3%), and BMI was recorded on the urine-test day for 6.1% (2285/37,443) of them.

XSL•FO
RenderX

Table 1. Baseline characteristics of the development cohort and the KNHANES[a] external validation cohort.

Characteristic	Development cohort (n=129,535)[b]	KNHANES adults (n=49,987)[c]
Demographics		
Years of testing	2010-2025	2016-2024 (survey years)
Age (years)		
Age, mean (SD)	49.0 (13.5)	52.5 (16.8)
19-39, n (%)	33,918 (26.2)	12,643 (25.3)
40-59, n (%)	66,426 (51.3)	18,198 (36.4)
60-79, n (%)	27,467 (21.2)	16,772 (33.6)
≥80[d], n (%)	1724 (1.3)	2374 (4.7)
Male sex, n (%)	67,443 (52.1)	22,281 (44.6)
BMI (kg/m^2), mean (SD)	23.8 (3.6)	24.0 (3.6)
BMI (≥25 kg/m^2), n (%)	43,610 (33.7)	17,910 (35.8)
BMI from a same-day health checkup[e], n (%)	105,704 (81.6)	N/A[f]
Urine dipstick		
Glucose (trace or above), n (%)	4077 (3.1)	2717 (5.4)
Protein (trace or above), n (%)	23,880 (18.4)	7140 (14.3)
Ketone (trace or above)[g], n (%)	35,157 (27.1)	1536 (3.1)
Bilirubin (1+ or above)[g,h], n (%)	4646 (3.6)	965 (1.9)
Occult blood (trace or above), n (%)	19,025 (14.7)	8451 (16.9)
Urobilinogen (above normal)[i], n (%)	141 (0.11)	430 (0.86)
Nitrite (positive), n (%)	1035 (0.80)	1188 (2.4)
pH, mean (SD)	5.81 (0.89)	5.88 (0.77)
Specific gravity[j], mean (SD)	1.021 (0.006)	1.019 (0.007)
Serum urate		
Serum urate (mg/dL)[k], mean (SD)	5.46 (1.52)	5.08 (1.39)
Serum urate (mg/dL), median (IQR)	5.30 (4.40-6.40)	4.90 (4.10-5.90)
Hyperuricemia (mg/dL), n (%)		
≥7.0 (primary)	21,124 (16.3)	5068 (10.1)[l]
≥9.0 (secondary)	2470 (1.9)	444 (0.89)
≥10.0 (exploratory)	736 (0.57)	103 (0.21)

[a]KNHANES: Korea National Health and Nutrition Examination Survey.

[b]Persons aged ≥19 years at their first urine test, with serum urate and BMI recorded on the day of the test and all 12 inputs available (1 record per person).

[c]Participants aged ≥19 years with all 12 inputs and serum urate available; dipstick results were coded using the label-matched mapping X1 (Table S2 in Multimedia Appendix 2).

[d]KNHANES age is top-coded at 80 years.

[e]BMI measured during a health checkup on the day of the urine test was the only marker of care setting in the data; for the other persons, BMI was calculated from height and weight recorded in the nursing record on that day.

[f]N/A: not applicable.

[g]Positivity in the development cohort decreased following a change in laboratory reporting in late 2021 (Table S15 in Multimedia Appendix 2).

[h]The development laboratory did not report a trace level; KNHANES "+" results were coded as 1+ (Table S2 in Multimedia Appendix 2).

[i]Above normal: 1+ or higher in the development data and "+" or higher in KNHANES, where normal and trace results were both coded as normal in the development data (Table S2 in Multimedia Appendix 2).

[j]Increments of 0.005 up to 1.030 in the development data and increments of 0.001 up to 1.050 in KNHANES (4.9% had values above 1.030).

XSL•FO
RenderX

[k]Results reported as "<1.0 mg/dL" were counted as 0.5 mg/dL (40 persons in the development cohort).

[l]Unweighted. Using the KNHANES examination weights, prevalence was 12% for serum urate levels ≥7.0 mg/dL, 1.08% for levels ≥9.0 mg/dL, and 0.24% for levels ≥10.0 mg/dL.

Discrimination and Calibration

The AUROC was 0.825 internally and 0.845 externally at ≥7.0 mg/dL and 0.859 and 0.860 at ≥9.0 mg/dL, and the external AUPRC was 0.373 and 0.070, 3.7 and 7.9 times the prevalence (Table 2 and Figure 2).

Internally, the O:E ratio was 1.00 at both thresholds; externally, the models overpredicted (O:E 0.73 at ≥7.0 mg/dL and 0.48 at ≥9.0 mg/dL), with negative calibration intercepts and slopes near 1 (Table 2 and Figure 3). In the highest decile of predicted risk, the mean prediction was 47.9% against an observed 39.2% at ≥7.0 mg/dL and 9.1% against 4.7% at ≥9.0 mg/dL (Table S11 in Multimedia Appendix 2). At ≥9.0 mg/dL, the external scaled Brier score was 0.002 (95% CI –0.015 to 0.017), no better than assigning every participant the prevalence. An intercept update estimated in one half of KNHANES restored the O:E ratio to 1.00 in the other half at both thresholds and raised the scaled Brier score to 0.175 and 0.032.

Isotonic calibration assigned a predicted risk of exactly 0 to 11.2% of internal test persons and 5.8% of KNHANES adults at ≥9.0 mg/dL, none of whom had an event. Platt and logistic calibrators fitted on the same fold learners produced no zeros and changed the AUROC by at most 0.0007, and flooring the isotonic outputs at 0.001 left the Brier score unchanged (Table S11 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 2. Discrimination and calibration of the models in the internal test set and in KNHANES[a] adults, with external calibration after an intercept-only update. The 95% CIs are percentile bootstrap intervals based on 1000 resamples. Scaled Brier score: 1 minus the Brier score divided by the Brier score of a constant prediction at the observed prevalence; values near 0 indicate performance no better than that of the constant prediction, and negative values indicate worse performance.

Measure	≥7.0 mg/dL (internal[b])	≥7.0 mg/dL (external[c])	≥9.0 mg/dL (internal[b])	≥9.0 mg/dL (external[c])
Cohort				
Persons, n	25,907	49,987	25,907	49,987
Events, n (%)	4225 (16.3)	5068 (10.1)	494 (1.9)	444 (0.89)
Discrimination				
AUROC[d] (95% CI)	0.825 (0.820 to 0.831)	0.845 (0.840 to 0.849)	0.859 (0.845 to 0.873)	0.860 (0.845 to 0.873)
AUPRC[e] (95% CI)	0.447 (0.432 to 0.463)	0.373 (0.359 to 0.387)	0.131 (0.109 to 0.161)	0.070 (0.056 to 0.089)
AUPRC relative to prevalence (95% CI)	2.7 (2.6 to 2.8)	3.7 (3.5 to 3.8)	6.9 (5.7 to 8.4)	7.9 (6.4 to 10.0)
Calibration				
Mean predicted risk, %	16.3	13.9	1.90	1.86
Calibration intercept (95% CI)	0.00 (–0.03 to 0.04)	–0.43 (–0.46 to –0.40)	0.00 (–0.09 to 0.09)	–0.78 (–0.88 to –0.69)
Calibration slope (95% CI)	0.96 (0.92 to 0.99)	1.07 (1.04 to 1.10)	1.01 (0.93 to 1.10)	1.10 (1.00 to 1.19)
O:E[f] ratio (95% CI)	1.00 (0.98 to 1.03)	0.73 (0.71 to 0.75)	1.00 (0.92 to 1.08)	0.48 (0.43 to 0.52)
Brier score (95% CI)	0.1096 (0.1072 to 0.1119)	0.0775 (0.0761 to 0.0791)	0.0175 (0.0161 to 0.0189)	0.0088 (0.0080 to 0.0095)
Scaled Brier score (95% CI)	0.197 (0.187 to 0.206)	0.149 (0.138 to 0.160)	0.063 (0.050 to 0.076)	0.002 (–0.015 to 0.017)
Predicted risk exactly 0, n (%); events[g]	38 (0.15); 0	26 (0.05); 0	2890 (11.2); 0	2895 (5.8); 0
External, after an intercept-only update[h]				
Intercept update, log-odds (95% range)	N/A[i]	–0.43 (–0.44 to –0.42)	N/A	–0.78 (–0.79 to –0.76)
O:E ratio (95% range)	N/A	1.00 (0.98 to 1.02)	N/A	1.00 (0.97 to 1.02)
Scaled Brier score (95% range)	N/A	0.175 (0.169 to 0.180)	N/A	0.032 (0.027 to 0.037)
Calibration slope if refitted (95% range)	N/A	1.07 (1.04 to 1.10)	N/A	1.10 (1.02 to 1.18)

[a]KNHANES: Korea National Health and Nutrition Examination Survey.

[b]Internal test set: 25,907 persons from the development cohort held out from training (1 record per person; the same persons were used for all thresholds).

[c]KNHANES participants from 2016 to 2024 aged ≥19 years; dipstick results were coded using the label-matched mapping X1 (Table S2 in Multimedia Appendix 2).

[d]AUROC: area under the receiver operating characteristic curve.

[e]AUPRC: area under the precision-recall curve.

[f]O:E: observed-to-expected.

[g]Isotonic calibration assigned a predicted risk of exactly 0 to some persons (Table S11 in Multimedia Appendix 2).

[h]Intercept update was fitted on 1 random half of KNHANES and evaluated on the other half (100 outcome-stratified splits; mean and 2.5th and 97.5th percentiles across splits).

[i]N/A: not applicable.

XSL•FO
RenderX

Figure 2. Receiver operating characteristic curves of the models in the internal test set (dashed line) and in KNHANES (Korea National Health and Nutrition Examination Survey) adults (solid line), showing (A) serum urate ≥7.0 mg/dL and (B) serum urate ≥9.0 mg/dL. Cases, persons, and area under the receiver operating characteristic curve (AUROC) with 95% CI are identified, and the dotted line indicates chance. y: years.

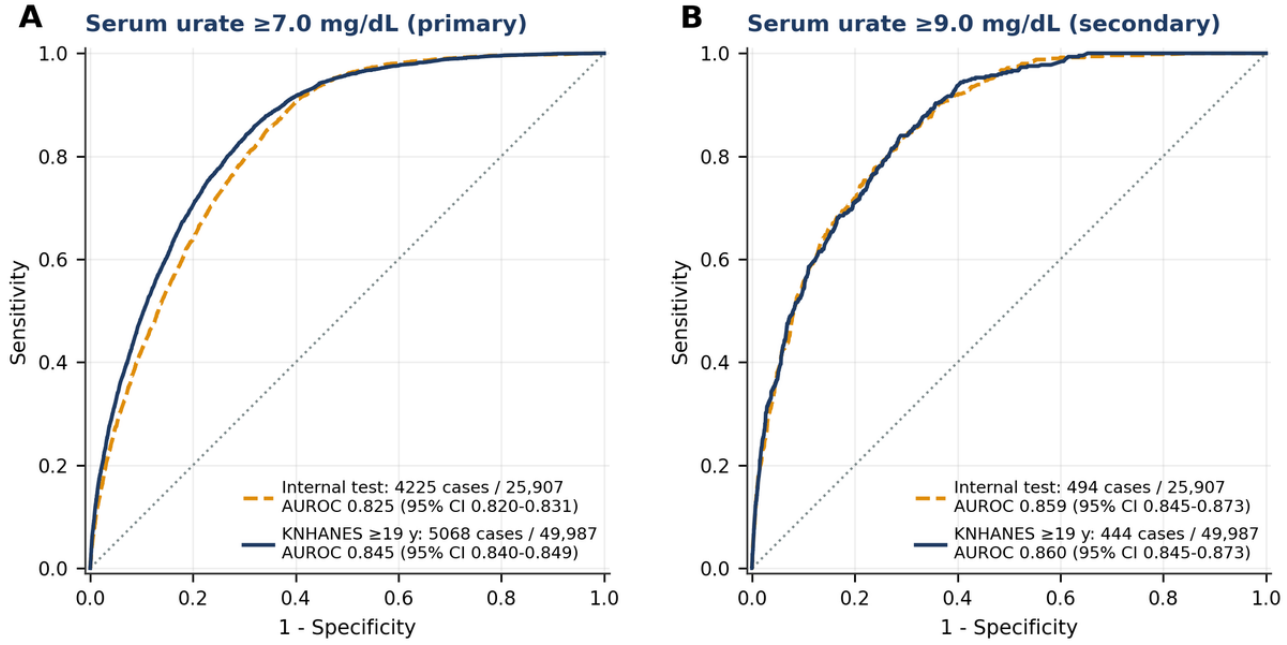
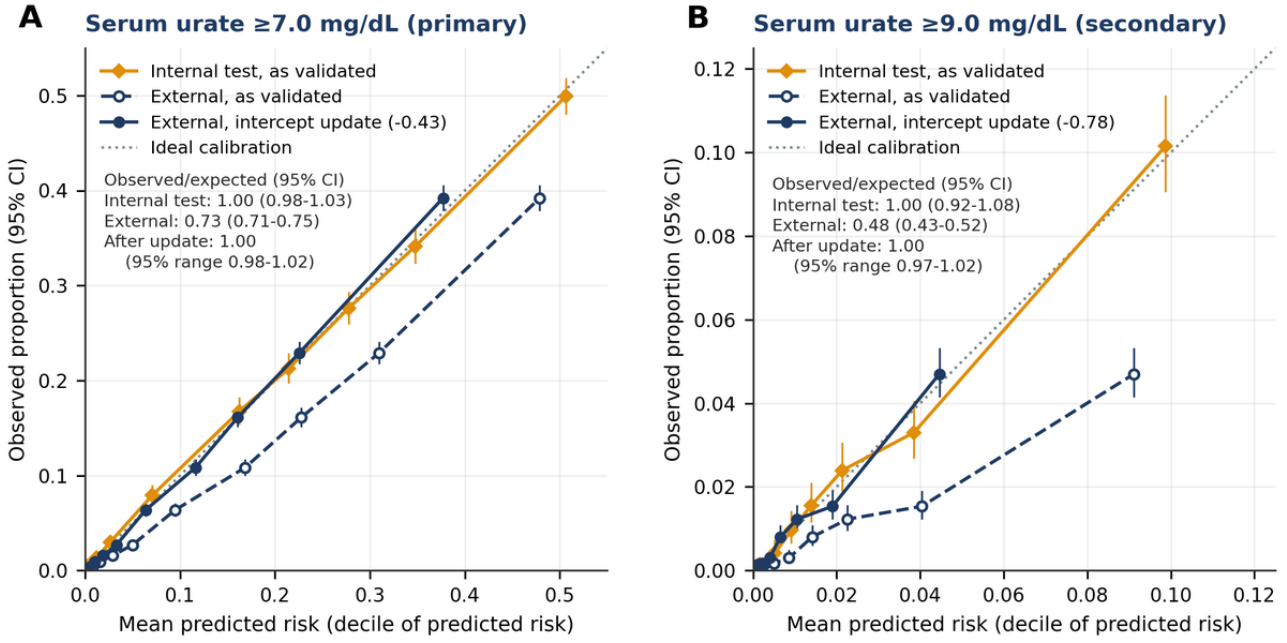


Figure 3. Calibration of the models in deciles of predicted risk for (A) serum urate ≥7.0 mg/dL and (B) serum urate ≥9.0 mg/dL. Diamonds represent the internal test set, open circles represent KNHANES (Korea National Health and Nutrition Examination Survey) adults as validated, and filled circles represent KNHANES adults after an intercept-only update (log-odds offsets of –0.43 and –0.78, the means of fits on 100 random halves of KNHANES). Observed proportions are presented with Wilson 95% CIs, and the dotted line indicates ideal calibration. The observed-to-expected (O:E) ratio after the update was evaluated in the held-out halves, with the mean and 2.5th to 97.5th percentiles calculated over the 100 splits. Ties at a predicted risk of 0 merge the lowest internal deciles at ≥9.0 mg/dL into 9 bins.



Clinical Utility

At ≥7.0 mg/dL, the 80%-sensitivity threshold chosen on the internal test set (0.195) flagged 29.8% of KNHANES adults, with a sensitivity of 77.1% (95% CI 75.9%-78.2%) and a PPV of 26.2% (Table 3). Of flagged adults, 97.4% were men, and sensitivity was 84.9% in men but 10.1% in women. Testing all men flagged 44.6% of adults with 89.6% sensitivity and 20.4% PPV. Sex-specific thresholds reached 86.3% sensitivity in women but flagged 47.0% of them at a PPV of 3.5%. The 80%-sensitivity threshold chosen on out-of-fold training predictions (0.200) gave a sensitivity of 75.8% (95% CI 74.6%-77.0%) in KNHANES, whereas thresholds chosen in one random half of KNHANES gave 80.1% in the other half (95% range over 100 splits, 77.8%-82.3%) (Table S12 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 3. External operating points of the models and of reference strategies in KNHANES[a] adults, overall and by sex. Thresholds (in parentheses) are predicted probabilities. KNHANES adults (n = 49,987): 22,281 men and 27,706 women, with 4,541 and 527 events, respectively, at ≥7.0 mg/dL and 399 and 45 events, respectively, at ≥9.0 mg/dL. Predictions were not recalibrated (after an intercept update; panel E in Table S12 in Multimedia Appendix 2); thresholds were fixed.

Strategy (threshold)	Flagged (men; women), %	Men among flagged, %	Sensitivity, % (95% CI)	Sensitivity (men; women), %	PPV[b], % (95% CI)	PPV (men; women), %
≥7.0 mg/dL (primary model)						
Ensemble model, internally derived thresholds[c]						
Youden index (0.177)	33.5 (72.8; 1.9)	96.9	81.5 (80.3-82.5)	89.5; 12.1	24.7 (24.0-25.3)	25.1; 12.3
Sensitivity 80% (0.195)	29.8 (65.2; 1.4)	97.4	77.1 (75.9-78.2)	84.9; 10.1	26.2 (25.5-26.9)	26.5; 13.8
Sensitivity 90% (0.152)	38.6 (83.1; 2.9)	95.9	86.8 (85.9-87.8)	94.8; 18.4	22.8 (22.2-23.4)	23.3; 12.1
Top 10% referred (0.365)[d]	10.0 (22.2; 0.19)	98.9	38.8 (37.6-40.0)	42.9; 2.8	39.3 (37.9-40.7)	39.4; 28.3
Sex-specific, sensitivity 80% (men 0.221; women 0.0204)[e]	51.6 (57.4; 47.0)	49.6	80.1 (79.0-81.1)	79.4; 86.3	15.7 (15.3-16.1)	28.2; 3.5
Reference strategies						
Test all men	44.6 (100; 0)	100	89.6 (88.8-90.4)	100; 0	20.4 (19.8-20.9)	20.4; N/A[f]
Age, sex, BMI model, sensitivity 80% (0.197)[g]	32.7 (73.0; 0.34)	99.43	76.0 (74.9-77.1)	84.7; 1.5	23.5 (22.9-24.2)	23.6; 8.6
Age, sex, BMI model, top 10% (0.363)	10.0 (22.4; 0)	100	36.5 (35.3-37.9)	40.8; 0	37.0 (35.8-38.5)	37.0; N/A
≥9.0 mg/dL (secondary model)						
Ensemble model, internally derived thresholds[c]						
Youden index (0.0178)	30.3 (61.2; 5.4)	90.1	84.0 (80.4-87.2)	89.2; 37.8	2.5 (2.2-2.7)	2.6; 1.1
Sensitivity 80% (0.0196)	28.3 (57.3; 4.9)	90.4	82.0 (78.2-85.4)	87.0; 37.8	2.6 (2.3-2.8)	2.7; 1.3
Sensitivity 90% (0.0122)	37.0 (73.2; 7.9)	88.2	90.5 (87.7-93.0)	94.2; 57.8	2.2 (2.0-2.4)	2.3; 1.2
Top 10% referred (0.0525)[d]	10.0 (21.3; 0.87)	95.2	52.9 (48.5-57.6)	57.4; 13.3	4.7 (4.2-5.3)	4.8; 2.5
Sex-specific, sensitivity 80% (men 0.0209; women 0.00740)[e]	31.3 (52.4; 14.3)	74.6	82.2 (78.8-85.6)	83.5; 71.1	2.3 (2.1-2.6)	2.9; 0.81
Reference strategies						
Test all men	44.6 (100; 0)	100	89.9 (86.7-92.6)	100; 0	1.8 (1.6-2.0)	1.8; N/A
Age, sex, BMI model, sensitivity 80% (0.0199)[g]	36.7 (75.7; 5.4)	91.9	81.5 (78.0-84.7)	88.2; 22.2	2.0 (1.8-2.2)	2.1; 0.67
Age, sex, BMI model, top 10% (0.0459)	10.0 (22.0; 0.36)	98.0	40.5 (35.7-45.0)	45.1; 0	3.6 (3.1-4.1)	3.7; 0

[a]KNHANES: Korea National Health and Nutrition Examination Survey.

[b]PPV: positive predictive value.

[c]Threshold selected on the internal test set (apparent performance on that set; panels A and C in Table S12 in Multimedia Appendix 2) and applied unchanged to KNHANES.

[d]Unlike the other rows in this group, the cut-off was determined separately in each cohort: the 10% of KNHANES adults with the highest predicted risk (internal top-10% cut-offs of 0.398 at ≥7.0 mg/dL and 0.0517 at ≥9.0 mg/dL); participants tied at the cut-off were flagged fractionally. 95% CIs were calculated as described in panel A in Table S12, in Multimedia Appendix 2, with the top 10% reselected in each resample (thresholds were fixed for the other rows).

[e]Separate thresholds were selected to achieve 80% sensitivity in men and women in the internal test set.

XSL•FO
RenderX

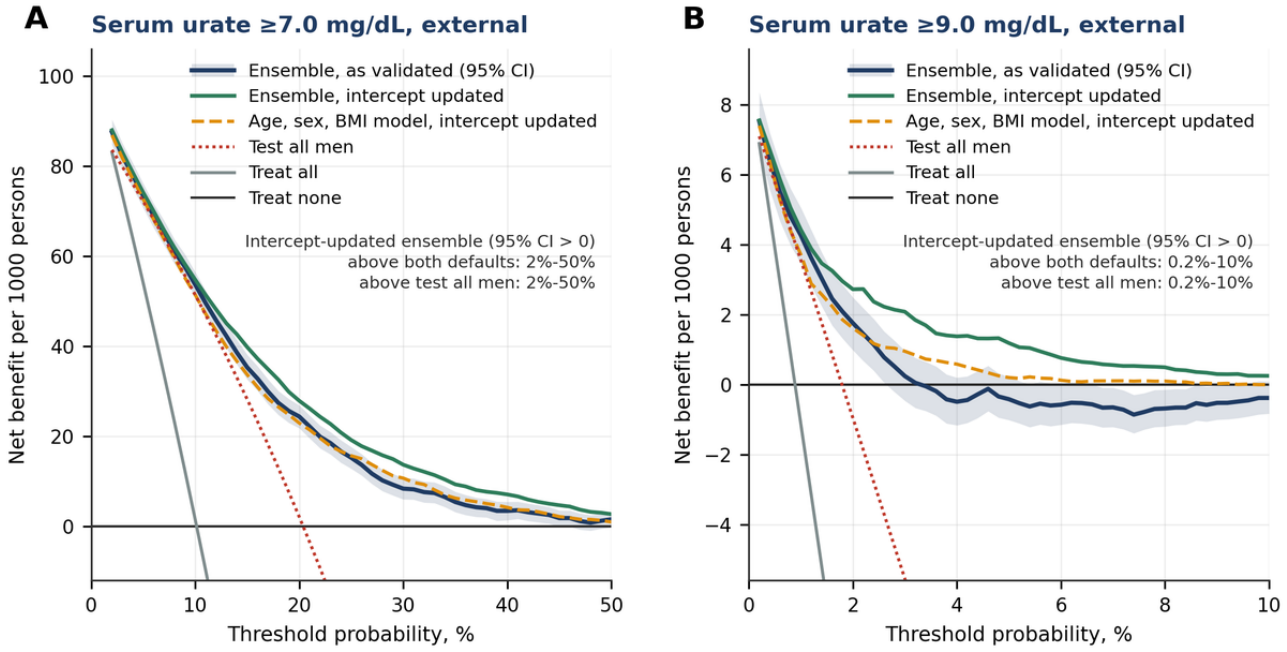
[f]N/A: not applicable.

[g]Logistic regression with splines for age and BMI and an age-by-sex interaction term, fitted using the training persons; thresholds were selected on the internal test set.

The net benefit of the model as validated exceeded that of testing all men at every threshold probability from 2% to 50% (lower 95% CI bound above 0 throughout): at 10%, it was 53.7 per 1000 adults against 51.4 for testing all men, and 54.8 after the intercept update (difference from testing all men 3.4 per 1000, 95% CI 2.6-4.1; Figure 4 and Table S13 in Multimedia Appendix 2). The net benefit of testing all men stayed above that of testing no one up to a threshold probability of 20% (2.1 per 1000) and fell below it from 21% (–3.5 per 1000).

At ≥9.0 mg/dL, the transferred 80%-sensitivity threshold (0.0196) gave a sensitivity of 82% (95% CI 78.2%-85.4%; 87% in men and 37.8%, 17 of 45, in women) and a PPV of 2.6%, flagging 28.3% (14,130/49,987) of adults. Net benefit exceeded both default strategies up to a threshold probability of 3.2% as validated and, after the intercept update, across the whole range examined (0.2%-10%). On the recalibrated scale, a risk threshold of 2% flagged 14% of adults at 61.7% sensitivity and 3.9% PPV (Table S12 in Multimedia Appendix 2).

Figure 4. External decision curves in KNHANES (Korea National Health and Nutrition Examination Survey) adults for (A) serum urate ≥7.0 mg/dL, with threshold probabilities of 2% to 50%, and (B) serum urate ≥9.0 mg/dL, with threshold probabilities of 0.2% to 10%. The curves show net benefit per 1000 adults for the ensemble as validated (95% bootstrap CI band) and after an intercept update cross-fitted in KNHANES (mean over 100 random splits), the age-sex-BMI spline logistic model after the same update, testing all men, and treating all or none. Annotations indicate the threshold ranges in which the updated ensemble exceeded both defaults and testing all men, with the lower 95% CI bound above 0.



Added Value of the Dipstick and Comparison With Simpler Models

In comparison models fitted on the training persons, age and sex alone reached an external AUROC of 0.781 to 0.799 at ≥7.0 mg/dL across model classes, age, sex, and BMI 0.814 to 0.823, and the dipstick fields alone 0.619 to 0.629 (Table 4 and Table S9 in Multimedia Appendix 2). Adding the dipstick fields to age, sex, and BMI increased the external AUROC by 0.017 to 0.024 at ≥7.0 mg/dL and 0.052 to 0.074 at ≥9.0 mg/dL across the 4 model classes, and the ensemble exceeded the best age-sex-BMI model (spline logistic regression) by 0.022 (95% CI 0.019-0.024) and 0.055 (95% CI 0.042-0.068). Dipstick fields with age and sex reached an external AUROC similar to that of age, sex, and BMI (0.820 vs 0.816 with LightGBM), and across 3 model classes the dipstick added 0.021 to 0.028 to age and sex at ≥7.0 mg/dL and 0.064 to 0.075 at ≥9.0 mg/dL. Within each sex, the increment over age, sex, and BMI at ≥7.0 mg/dL was larger: 0.038 to 0.049 in men and 0.034 to 0.074 in women with LightGBM and spline logistic regression (Table S9 in Multimedia Appendix 2). With referral of the top 10% (fixed-configuration LightGBM), the dipstick added 3.1 (95% CI 2.1-4.0) cases detected per 1000 adults screened at ≥7.0 mg/dL (38.6 vs 35.5 with the same number of tests; 0.8-4.4 for referral of the top 1%-20%) and 1.6 (95% CI 1.1-2.1) at ≥9.0 mg/dL (4.3 vs 2.8; Table S13 in Multimedia Appendix 2).

At ≥7.0 mg/dL, the ensemble's AUROC was within 0.0011 of that of a grid-tuned single LightGBM and of its own LightGBM member (a comparison that isolates the addition of XGBoost and CatBoost); at ≥9.0 mg/dL, it exceeded them by 0.0020 to 0.0024 internally and 0.006 to 0.008 externally (Table 4). Spline logistic regression with the same 12 inputs was 0.008 (internal) and 0.005 (external) below the ensemble at ≥7.0 mg/dL and not significantly different externally at ≥9.0 mg/dL (0.0025, 95% CI -0.0052 to 0.0094; Table S10 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 4. Added value of the urine dipstick over age and sex and over age, sex and BMI, and comparison of the ensemble with simpler models (AUROC[a]; 2F, 3F, 9F, 11F, and 12F indicate input sets containing 2, 3, 9, 11, and 12 inputs, respectively).

Model and input set	≥7.0 mg/dL (internal)	≥7.0 mg/dL (external[b])	≥9.0 mg/dL (internal)	≥9.0 mg/dL (external[b])
AUROC by input set[c] (95% CI)				
LightGBM[d], fixed configuration				
Age and sex	0.778 (0.771 to 0.785)	0.798 (0.793 to 0.803)	0.760 (0.743 to 0.776)	0.757 (0.740 to 0.772)
Age, sex, and BMI	0.800 (0.794 to 0.806)	0.816 (0.811 to 0.822)	0.780 (0.763 to 0.797)	0.768 (0.748 to 0.785)
Dipstick only	0.624 (0.615 to 0.633)	0.619 (0.611 to 0.627)	0.694 (0.669 to 0.719)	0.656 (0.631 to 0.681)
Dipstick, age, and sex	0.806 (0.800 to 0.812)	0.820 (0.815 to 0.825)	0.842 (0.828 to 0.857)	0.821 (0.804 to 0.837)
Dipstick, age, sex, and BMI	0.823 (0.818 to 0.829)	0.841 (0.836 to 0.846)	0.849 (0.834 to 0.864)	0.842 (0.827 to 0.857)
Logistic regression, age and BMI splines, age-by-sex term				
Age and sex	0.778 (0.771 to 0.785)	0.799 (0.794 to 0.805)	0.762 (0.745 to 0.778)	0.777 (0.759 to 0.794)
Age, sex, and BMI	0.802 (0.795 to 0.808)	0.823 (0.818 to 0.828)	0.792 (0.775 to 0.808)	0.805 (0.787 to 0.821)
Dipstick, age, and sex	0.799 (0.793 to 0.805)	0.821 (0.815 to 0.826)	0.840 (0.826 to 0.854)	0.843 (0.828 to 0.857)
Dipstick, age, sex, and BMI	0.818 (0.812 to 0.823)	0.839 (0.834 to 0.844)	0.851 (0.837 to 0.865)	0.857 (0.843 to 0.870)
Increments (AUROC difference, 95% CI)				
Dipstick added to age, sex, BMI (12F – 3F)				
LightGBM (fixed)	+0.023 (0.020 to 0.027)	+0.024 (0.021 to 0.028)	+0.069 (0.054 to 0.084)	+0.074 (0.059 to 0.091)
LightGBM (tuned[e])	+0.023 (0.020 to 0.027)	+0.021 (0.019 to 0.024)	+0.068 (0.054 to 0.084)	+0.059 (0.046 to 0.074)
Logistic regression (main effects)	+0.017 (0.014 to 0.020)	+0.019 (0.016 to 0.021)	+0.067 (0.052 to 0.082)	+0.053 (0.041 to 0.066)
Logistic regression (splines)	+0.016 (0.013 to 0.019)	+0.017 (0.015 to 0.019)	+0.059 (0.046 to 0.072)	+0.052 (0.042 to 0.063)
Dipstick added to age and sex (11F – 2F)				
LightGBM (fixed)	+0.028 (0.023 to 0.032)	+0.022 (0.019 to 0.026)	+0.083 (0.064 to 0.102)	+0.064 (0.048 to 0.081)
Logistic regression (main effects)	+0.023 (0.019 to 0.028)	+0.028 (0.025 to 0.032)	+0.097 (0.077 to 0.117)	+0.075 (0.060 to 0.091)
Logistic regression (splines)	+0.021 (0.017 to 0.024)	+0.021 (0.019 to 0.024)	+0.078 (0.062 to 0.093)	+0.066 (0.053 to 0.080)
BMI added to age and sex (3F – 2F)				
LightGBM (fixed)	+0.022 (0.018 to 0.026)	+0.018 (0.015 to 0.021)	+0.021 (0.006 to 0.036)	+0.011 (–0.003 to 0.025)
Logistic regression (splines)	+0.024 (0.020 to 0.027)	+0.023 (0.020 to 0.026)	+0.029 (0.016 to 0.042)	+0.028 (0.016 to 0.040)
Ensemble vs simpler 12-input models[f]				
Ensemble, AUROC (95% CI)	0.825 (0.820 to 0.831)	0.845 (0.840 to 0.849)	0.859 (0.845 to 0.873)	0.860 (0.845 to 0.873)
Minus main-effects logistic regression	+0.013 (0.010 to 0.016)	+0.012 (0.010 to 0.015)	+0.026 (0.018 to 0.035)	+0.014 (0.003 to 0.025)
Minus spline logistic regression	+0.008 (0.006 to 0.010)	+0.005 (0.004 to 0.007)	+0.008 (0.001 to 0.015)	+0.0025 (–0.0052 to 0.0094)
Minus LightGBM, grid-tuned	–0.00002 (–0.0004 to 0.0004)	+0.0009 (0.0004 to 0.0013)	+0.0020 (–0.0001 to 0.0040)	+0.006 (0.002 to 0.009)
Minus LightGBM, ensemble member configuration[g]	+0.00003 (–0.0003 to 0.0004)	+0.0011 (0.0007 to 0.0015)	+0.0024 (0.0004 to 0.0044)	+0.008 (0.004 to 0.011)

[a]AUROC: area under the receiver operating characteristic curve.

[b]KNHANES adults, with dipstick results coded using the label-matched mapping X1 for all models; with the alternative codings M1 and M2, the external

XSL•FO
RenderX

AUROCs of models with dipstick inputs changed by no more than 0.0011 (Table S4 in Multimedia Appendix 2).

[c]Comparison models were refitted for this analysis using the training persons from the development cohort (n = 103,628) and evaluated on the internal test persons (n=25,907) and KNHANES adults; differences were calculated using paired bootstrap resamples.

[d]LightGBM: Light Gradient Boosting Machine.

[e]LightGBM was tuned separately for each input set (12 configurations were compared using 3-fold cross-validation on the training persons).

[f]Comparison models with the same 12 inputs, fitted using the same 103,628 training persons as the ensemble and evaluated on the same persons; differences represent the ensemble minus the comparison model.

[g]The LightGBM member of the ensemble with the same hyperparameters and the same 5-fold isotonic calibration (its fold learners are identical to the ensemble's LightGBM members), so the difference is the contribution of adding XGBoost and CatBoost.

Missing Data

Among the 534,887 adults whose first urine test had a blood test within 14 days, BMI was not recorded on the urine-test day for 74.8% (400,043/534,887), and serum urate was not measured on that day for 17.3% (92,711/534,887); each dipstick field was missing for 0.01% or less (Table S5 in Multimedia Appendix 2). The Little test rejected missingness completely at random, also in each of 10 random subsamples of every size from 1000 to 100,000 persons. The 312,613 persons excluded only because BMI was missing were older than included persons (mean age 55.7, SD 16.0 years vs 49.0, SD 13.5 years), had more alkaline and more dilute urine, and less often had serum urate ≥7.0 mg/dL (39,028/312,613, 12.5% vs 21,124/129,535, 16.3%). In these persons, the models with BMI imputed from the other inputs and serum urate had an AUROC of 0.815 at ≥7.0 mg/dL and 0.853 at ≥9.0 mg/dL (O:E 0.91 and 0.93; calibration slopes 0.90 and 1.01); with BMI imputed from age and sex only, the AUROC was 0.778 and 0.805 and the calibration slopes 0.74 and 0.78 (Table 5 and Table S5 in Multimedia Appendix 2). In KNHANES, the 3383 (6.1%) adults without serum urate were similar in age and sex to those with it (Table S1 in Multimedia Appendix 2).

XSL•FO
RenderX

Table 5. Sensitivity analyses of the models: discrimination (AUROC[a]) and calibration (O:E[b] ratio). Point estimates (95% CIs are provided in Multimedia Appendix 2); O:E values below 1 indicate overprediction. Rows without n are based on the full internal test set or all KNHANES adults.

Analysis	n (events at ≥7.0 mg/dL)	≥7.0 mg/dL, AUROC	≥7.0 mg/dL, O:E ratio	≥9.0 mg/dL, AUROC	≥9.0 mg/dL, O:E ratio
Reference					
Internal test set	25,907 (4225)	0.825	1.00	0.859	1.00
KNHANES adults	49,987 (5068)	0.845	0.73	0.860	0.48
Setting and selection (internal test set unless stated)					
BMI from a same-day health checkup[c]	21,143 (3521)	0.826	1.01	0.858	0.95
BMI from the nursing record	4764 (704)	0.819	0.95	0.849	1.15
Weighted for urate measurement[d]	N/A[e]	0.828	1.00	0.861	1.01
Same, measurement model among persons with BMI	N/A	0.826	1.00	0.859	1.01
Excluded for missing BMI, BMI imputed (full model)[f]	312,613 (39,028)	0.815	0.91	0.853	0.93
Same, BMI imputed from age and sex	312,613 (39,028)	0.778	0.91	0.805	0.93
Time and population					
Urine tests in 2010-2021	20,384 (3482)	0.824	1.05	0.858	1.10
Urine tests in 2022-2025	5523 (743)	0.832	0.84	0.872	0.71
Temporal holdout, refitted[g]	27,778 (3717)	0.840	0.79	0.866	0.61
KNHANES single survey years (9), range	N/A	0.835-0.860	0.66-0.88	0.823-0.888	0.39-0.57
KNHANES with survey weights[h]	N/A	0.840	0.74	0.855	0.51
KNHANES input coding[i]					
Bilirubin coded higher, urobilinogen lower (M1)	N/A	0.845	0.73	0.860	0.46
Urobilinogen normal coded 0 (M2)	N/A	0.845	0.73	0.860	0.46
As M1, urobilinogen normal coded 0.5 (M3)	N/A	0.845	0.73	0.860	0.48
Specific gravity rounded to the development grid (X2)	N/A	0.844	0.74	0.856	0.50
Adults with specific gravity ≤1.030	47,532 (4855)	0.846	0.74	0.858	0.48
Ketone set to 0	N/A	0.844	0.73	0.858	0.48
Model form (refitted; KNHANES adults)					
Continuous urate model, constant variance	N/A	0.839	0.83	0.837	0.57
Logistic instead of isotonic calibration[j]	N/A	0.845	0.73	0.860	0.48

[a]AUROC: area under the receiver operating characteristic curve.

[b]O:E: observed-to-expected.

[c]Internal test persons stratified by the source of BMI recorded on the day of the urine test: a health checkup on that day or, otherwise, height and weight recorded in the nursing record (Table S6 in Multimedia Appendix 2).

[d]AUROC and O:E, with each internal test person weighted by the inverse probability of serum urate measurement on the day of the urine test. The measurement model was fitted among all adults with complete dipstick results, age, and sex (primary analysis) or among those with complete dipstick results, age, sex, and BMI (Table S7 in Multimedia Appendix 2).

[e]N/A: not applicable.

[f]Models applied to persons excluded because BMI was missing on the day of the urine test, with BMI multiply imputed (20 imputations) using age, sex, dipstick fields, and serum urate (full model) or age and sex only.

[g]The ensemble configuration, with hyperparameters unchanged, was refitted using persons tested from 2010 to 2021 and evaluated on those tested from 2022 to 2025.

[h]KNHANES examination weights, without readjustment for complete-case exclusion.

[i]Alternative codings of KNHANES dipstick results, each compared with the primary label-matched coding (X1; Tables S2 and S4 in Multimedia Appendix 2).

[j]The ensemble's fold learners with a logistic calibrator fitted to the log-odds of their soft-vote probabilities using the same calibration folds, instead of

XSL•FO
RenderX

isotonic regression (Table S11 in Multimedia Appendix 2).

Verification Bias and Setting

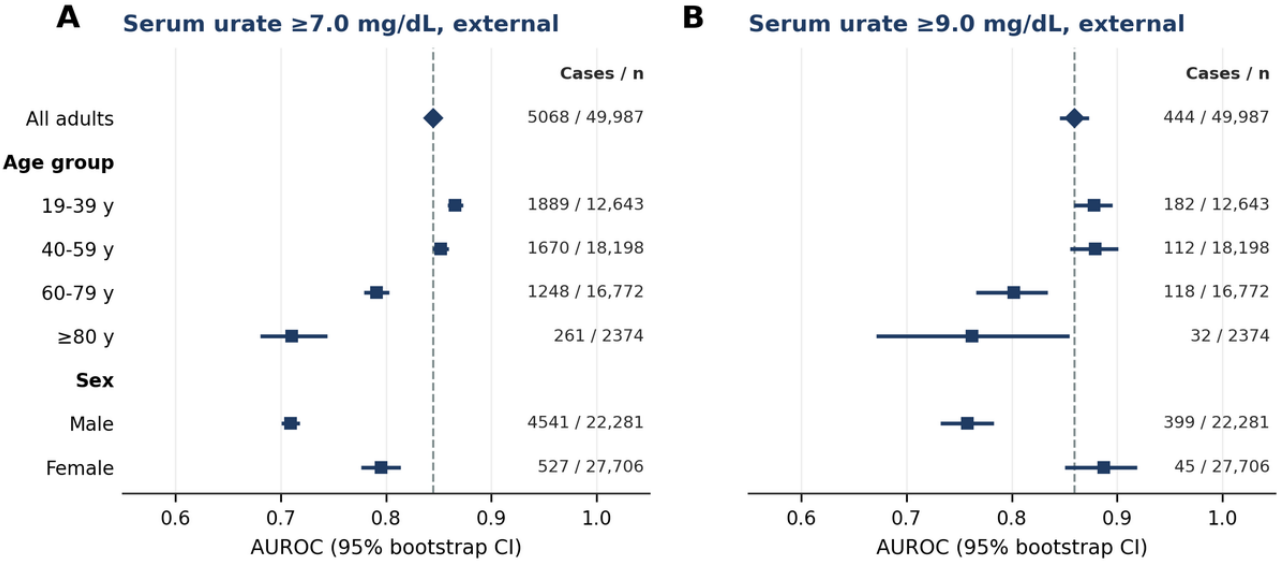
Whether serum urate was measured on the urine-test day depended on dipstick results, age, sex, and calendar year, but the measurement model had an AUROC of 0.615 (0.741 among persons with BMI recorded on that day; Table S7 in Multimedia Appendix 2). Reweighting by its inverse changed the internal AUROC at ≥7.0 mg/dL by +0.0023 (95% CI 0.0019-0.0028) with the primary measurement model and by +0.0002 with the model fitted among persons with BMI, and the O:E ratio by 0.01 or less (Table 5). Internally, the AUROC with and without a same-day health checkup (BMI from the nursing record) was 0.826 and 0.819 at ≥7.0 mg/dL and 0.858 and 0.849 at ≥9.0 mg/dL (O:E 0.95-1.15). Persons without a same-day checkup were older, and discrimination was lower in older persons (Table S17 in Multimedia Appendix 2). Within age bands, the AUROC at ≥7.0 mg/dL with checkup BMI minus that with nursing-record BMI was –0.021 (95% CI –0.040 to –0.002; *P*=.03) at 19 to 64 years and +0.023 (95% CI –0.020 to 0.065; *P*=.30) at 65 years or older (Table S6 in Multimedia Appendix 2).

Subgroups, Nested Thresholds, and Continuous-Urate Model

Discrimination declined with age: at ≥7.0 mg/dL, the external AUROC was 0.866 at 19 to 39 years and 0.711 at ≥80 years (Figure 5; Table S17 in Multimedia Appendix 2). In men, it was 0.710 at ≥7.0 mg/dL and 0.758 at ≥9.0 mg/dL, and in women 0.795 and 0.887 (95% CI 0.850-0.919; 45 events). The female-male difference was 0.086 (95% CI 0.066-0.106) at ≥7.0 mg/dL and 0.129 (95% CI 0.087-0.171) at ≥9.0 mg/dL, and 0.057 (95% CI 0.021-0.092) at ≥7.0 mg/dL among adults aged 40 to 69 years (Table S17 in Multimedia Appendix 2).

The models assigned a higher probability to ≥9.0 than to ≥7.0 mg/dL for 3 of 49,987 KNHANES adults and 1 of 25,907 internal test persons, each with a predicted risk of exactly 0 at ≥7.0 mg/dL and a difference of 0.0002; separately fitted binary LightGBM models did so for 180 KNHANES adults (largest difference 0.24), and the continuous-urate model, by construction, for none (Table S18 in Multimedia Appendix 2; comparisons with the exploratory ≥10.0 mg/dL model are in Table S22). The constant-variance continuous-urate model had an external AUROC of 0.839 at ≥7.0 mg/dL and 0.837 at ≥9.0 mg/dL, a difference of –0.0022 (95% CI –0.0037 to –0.0005) and –0.006 (95% CI –0.018 to 0.007) from binary LightGBM models with the same configuration and training persons, whose AUROCs were in turn 0.0038 and 0.017 below those of the ensembles. The continuous model's mean predicted urate equaled the observed mean internally (5.47 mg/dL) but exceeded it in KNHANES (5.28 vs 5.08 mg/dL; Table S18 in Multimedia Appendix 2).

Figure 5. Discrimination of the models by age group and sex in KNHANES (Korea National Health and Nutrition Examination Survey) adults for (A) serum urate ≥7.0 mg/dL and (B) serum urate ≥9.0 mg/dL. Squares represent subgroup area under the receiver operating characteristic curve (AUROC) with 95% CIs (bootstrap within subgroup), while the diamond and dashed line represent all adults. The numbers of cases and adults are shown on the right. Sex, the input with the largest importance, is constant within each sex, and the female estimate at ≥9.0 mg/dL is based on 45 events. Sex-specific sensitivity and positive predictive value (PPV) are provided in Table 3. y: years.



Sensitivity Analyses

Table 5 summarizes the sensitivity analyses. Alternative KNHANES codings changed the external AUROC by less than 0.005 and the O:E ratio by 0.02 or less, and survey weights lowered the AUROC by 0.005 and 0.0048 and raised the O:E ratio to 0.74 and 0.51. The O:E ratio differed more by period of the urine test: internally, it was 1.05 for urine tests in 2010-2021 and 0.84 (95% CI 0.79-0.89) in 2022-2025 at ≥7.0 mg/dL, and 1.10 and 0.71 (95% CI 0.57-0.85) at ≥9.0 mg/dL (Table S14 in Multimedia Appendix 2).

Checkup and nursing-record BMI of the same day agreed within 0.5 kg/m^2 for 99.94% (n=69,575) of the 69,617 adults with both; checkup weight showed no terminal-digit preference (*P*=.56) and checkup height little (10.6% ending in 0; *P*<.001), whereas the nursing-record weight and height used for 23,831 persons

XSL•FO
RenderX

ended in 0 in 19.6% (n=4660) and 19% (n=4529; *P*<.001 for both; Table S6 in Multimedia Appendix 2).

Rounding specific gravity to the development grid (X2) lowered the AUROC at ≥9.0 mg/dL by 0.0041 (95% CI 0.0017-0.0068); the other codings changed it by 0.0002 or less (Table S4 in Multimedia Appendix 2). The codings changed the AUROCs of the comparison models little (Table 4), but under M1 and M2 the main-effects logistic model with all 12 inputs predicted a mean risk of 19.8% and 19.7% at ≥7.0 mg/dL, against 13.2% under X1 and 10.1% observed (Table S4 in Multimedia Appendix 2). Among the inputs, the population stability index between the development cohort and KNHANES was largest for ketone (0.60) and specific gravity (0.49; 0.18 with specific gravity rounded to the development grid; Table S3 in Multimedia Appendix 2). Setting every KNHANES ketone result to negative left the O:E ratio at 0.73 and 0.48 (Table S16 in Multimedia Appendix 2).

With the reporting change at the development laboratory around December 2021, trace ketone results almost disappeared (1840/9039, 20.4% of persons tested in 2021 and 18/8798, 0.20% in 2022), as did bilirubin 1+ results (360/9039, 4% and 12/8798, 0.14%), whereas ketone 1+ or higher changed little (992/9039, 11% and 993/8798, 11.3%; Table S15 and Figure S1 in Multimedia Appendix 2). In the temporal holdout, an ensemble with the same configuration fitted on persons tested in 2010-2021 had an AUROC of 0.840 and 0.866 in persons tested in 2022-2025 and an O:E ratio of 0.79 and 0.61 (Table S14 in Multimedia Appendix 2). In KNHANES, the O:E ratio ranged from 0.66 to 0.88 at ≥7.0 mg/dL and from 0.39 to 0.57 at ≥9.0 mg/dL across survey years, without a monotone trend (Table S14 in Multimedia Appendix 2).

Interpretation of the Ensemble

In KNHANES, permutation importance ranked male sex first at both thresholds (AUROC drop 0.237 and 0.205), followed by BMI, age, pH, and specific gravity at ≥7.0 mg/dL and by BMI, specific gravity, age, and pH at ≥9.0 mg/dL; ensemble SHAP values identified the same 5 leading inputs. Permuting nitrite, bilirubin, occult blood, or urobilinogen changed the AUROC by 0.0005 or less (Table S19 and Figure S2 in Multimedia Appendix 2). Predicted risk fell with urine pH, rose with BMI, and fell with specific gravity above 1.015 in both sexes, whereas in the development cohort the crude association of specific gravity with serum urate ≥7.0 mg/dL was positive (odds ratio [OR] per 0.010, 1.12) and reversed after adjustment for sex (0.92) and for age, sex, and BMI (0.85). Between 40 and 70 years, predicted risk at ≥7.0 mg/dL fell by 43% in men and rose by 104% in women (Table S20 and Figure S3 in Multimedia Appendix 2).

Discussion

Principal Results

Routine dipstick results added a small increment to age, sex, and BMI for serum urate ≥7.0 mg/dL (external AUROC +0.017 to +0.024 across model classes), about 3 additional cases per 1000 adults screened with top-10% referral, which on its own is unlikely to justify implementation. Discrimination was similar in the hospital test set and the national survey sample (AUROC 0.825 and 0.845), but predicted risks were too high externally (O:E 0.73) and required local recalibration. At the thresholds tested, the models worked mainly as a selection rule for men: almost all flagged adults were men, most female cases were missed, and at threshold probabilities up to 10%, testing all men achieved nearly the same net benefit. Neither cohort represented encounters with urine but no blood, the setting that motivated the models.

The similarity of the internal and external AUROCs shows that the ranking transferred, not that the models generalize. A slope near 1 with a negative intercept means that the outcome was more frequent at the same predictor values in the development data than in KNHANES. Case mix did not explain it, because the mean predicted risk in KNHANES was only slightly lower than internally (13.9% vs 16.3% at ≥7.0 mg/dL and 1.86% vs 1.90% at ≥9.0 mg/dL), whereas the observed prevalence was much lower (10.1% vs 16.3% and 0.89% vs 1.9%) [42]. Input coding explained little: alternative KNHANES codings moved the O:E ratio by 0.02 or less, and setting every KNHANES ketone result to negative did not change it. Nor was the overprediction specific to the ensemble or the dipstick: LightGBM comparison models fitted on the same training persons over-predicted even more without the dipstick fields (O:E 0.66 vs 0.74 at ≥7.0 mg/dL and 0.33 vs 0.49 at ≥9.0 mg/dL; Table S13 in Multimedia Appendix 2). Candidate explanations, which these data cannot rank, are selection of hospital patients for urate testing on factors not recorded in the data, other differences between the populations beyond the measured inputs, differences between the hospital and KNHANES serum urate assays and their calibration, and a secular decline in prevalence at the hospital, visible in the lower O:E ratio for internal test persons tested in 2022-2025. As validated, the ≥9.0 mg/dL model was no more accurate in KNHANES than assigning every adult the observed prevalence (scaled Brier score 0.002). Because the miscalibration lay mainly in the intercept, an offset estimated in one half of KNHANES restored the O:E ratio to 1.00 in the other half; this is the simplest form of model updating [33], but it requires outcome data from the new setting and should precede any use there [30,32].

The intended use needs to be stated precisely. Where blood is drawn, serum urate should simply be measured, and a model has nothing to add. The motivating setting is encounters in which urine is tested without blood; at the development hospital these were 6.5% (37,443/572,330) of adults' first urine tests, 68.3% (25,559/37,443) of them in women, and BMI was recorded on the day of the urine test for only 6.1% (2285/37,443) of them. A comparison model without BMI (dipstick fields, age, and sex) discriminated about as well as one with age, sex, and BMI (external AUROC 0.820 vs 0.816 at ≥7.0 mg/dL with LightGBM), which makes it a candidate for that setting, but neither it nor the models have been evaluated there. After recalibration, the ≥9.0 mg/dL model's net benefit exceeded the default strategies across the examined range, so the narrow range as validated (0.2%-3.2%) reflected miscalibration rather than weak discrimination; its PPV nevertheless stays low. Because guidelines do not recommend urate-lowering therapy for asymptomatic hyperuricemia [8], the benefit of detection

XSL•FO
RenderX

lies in lifestyle counseling and attention to cardiovascular and renal risk.

The effects of the inputs were consistent with urate physiology: predicted risk fell with urine pH and rose with BMI [11,12,14], and the effect of age was opposite in men and women, consistent with the rise in serum urate after menopause [2]. The models do not treat concentrated urine as a risk factor; the crude positive association between specific gravity and serum urate reflects confounding by sex.

Comparison With Prior Work

Previous hyperuricemia models have mostly used blood or genetic inputs: machine learning on health checkup results in 38,001 Korean adults reached an AUROC of up to 0.775 [43], and a multimodal model reached 0.836 internally and 0.779 externally [44]. In our data, age, sex, and BMI alone reached an external AUROC of 0.814 to 0.823 at ≥7.0 mg/dL, so for models that include these variables a large part of the discrimination reflects demographics and body size, consistent with the clustering of gout with obesity and renal dysfunction [45]. We previously showed that machine learning improves chronic kidney disease detection from dipstick results [46]; for serum urate, the dipstick-specific gain is modest. The ensemble offered little over a single grid-tuned LightGBM model or a transparent spline logistic model with the same inputs (external AUROC differences 0.0009 and 0.005 at ≥7.0 mg/dL and 0.006 and 0.0025 at ≥9.0 mg/dL, the last not significant), so its additional complexity is not justified by its performance. We report it because its development procedure was fixed before evaluation, and replacing it after validation would have been post hoc model selection. For local updating, we would start from the 12-input spline logistic model, whose calibration is easier to inspect and which produces no predictions of exactly 0, although its predicted risks, unlike the ensemble's, shifted under alternative KNHANES codings and it discriminated less well in KNHANES women (AUROC 0.762 vs 0.795 for the ensemble at ≥7.0 mg/dL; Tables S4 and S9 in Multimedia Appendix 2).

The present models are cross-sectional: from 1 dipstick test they estimate the probability that serum urate measured on the same day is elevated. Longitudinal models that use full electronic health records [47], repeated measurements [48], or serum urate trajectories to predict incident hyperuricemia and gout [49] answer a different question. This study used each person's first urine test, so repeated measurements were not modeled; modeling them would require a different design and its own validation.

Implications for Implementation

Any use would require automatic transfer of the dipstick fields from the laboratory system; height and weight measured at the encounter (the models should not be used with imputed BMI; Table 5); a local intercept update, which needs outcome data from the intended setting; a sex-aware alert threshold chosen for the intended testing burden; and monitoring of input distributions, because the reporting change at the development laboratory around December 2021 showed how quickly dipstick coding can shift. If the ensemble is used, its isotonic outputs of exactly 0 should be floored or replaced by a logistic calibrator fitted on the same fold learners, which gave practically identical discrimination.

Limitations

First, the outcome is a laboratory abnormality, not gout. Alcohol intake and diet were not used as inputs because they are not routinely available at the point of care (alcohol was recorded only for health-checkup examinees), and medications were not linked.

Second, the development data come from 1 hospital, and requiring serum urate and BMI from the day of each person's first urine test selected patients in ways we cannot fully characterize. BMI was recorded on the urine-test day mainly at health checkups, so the development cohort consists mostly of health-checkup examinees (105,704/129,535, 81.6%) and of patients whose height and weight were entered in the nursing record. Neither cohort represents urine-only encounters, and no prospective, workflow, or cost-effectiveness evaluation was performed; evaluating a model in that setting requires a prospective study in which serum urate is measured in a sample of urine-only encounters, with the costs of recalling and testing patients weighed against the cases detected. Serum urate itself was measured selectively, and reweighting for the recorded determinants of its measurement cannot correct for unrecorded ones.

Third, BMI was missing on the urine-test day for 312,635 of the 442,176 adults with a same-day serum urate, and the data were not missing completely at random. Complete-case analysis is primary because the models require BMI measured at the encounter; in the adults excluded for missing BMI, performance with imputed BMI was lower, particularly with BMI imputed from age and sex only (Table 5). An imputation model using age, sex, and dipstick results without urate, the realistic point-of-care case, was not evaluated. Checkup BMI was measured, but the nursing-record height and weight used for 23,831 persons showed terminal-digit preference, so their measurement cannot be confirmed.

Fourth, hyperparameters were tuned for each learner separately, not for the calibrated ensemble; the equal voting weights were not tuned, and the search budget was small. Isotonic calibration, specified before the models were fitted, assigned a predicted risk of exactly 0 at ≥9.0 mg/dL to 11.2% (2890/25,907) of internal test persons and 5.8% (2895/49,987) of KNHANES adults. None of these persons had an event, but such values are implausible, the calibration slope depends on how they are handled (internally, 1.01 as fitted and 1.05 with a floor of 0.001), and the only reversals of threshold order between ≥7.0 and ≥9.0 mg/dL (a higher predicted probability for ≥9.0 mg/dL in 3 KNHANES adults and 1 internal test person) occurred where the ≥7.0 mg/dL prediction was exactly 0.

Fifth, the KNHANES coding, which matches grade labels and was fixed before the models were fitted, has not been validated by a crosswalk or paired-strip study. The alternative codings differed only for bilirubin, urobilinogen, and specific gravity; for protein, glucose, ketone, occult blood, nitrite, and pH, every coding assumed that results with the same label or value are

XSL•FO
RenderX

equivalent. The dipstick's increment over age, sex, and BMI was similar internally and in KNHANES (Table 4), which suggests that the information in the grades transferred but does not establish their equivalence; a paired-strip study would be needed before the models are used with another laboratory's strips. Specific gravity is measured by refractometry in KNHANES but probably by reagent pad at the hospital, and dipstick positivity varied between survey years. A reporting change at the development laboratory around December 2021 (transition from July 2021 to January 2022) almost removed trace ketone and bilirubin 1+ results; the models were trained on persons tested under both conventions, and the temporal holdout could not separate the change from the passage of time.

Sixth, internal operating points chosen on the test set are apparent performance, and at ≥7.0 mg/dL the 80%-sensitivity rule fell short of its target in KNHANES whether the threshold was derived in the internal test set (sensitivity 77.1%) or on out-of-fold training predictions (75.8%). With the internally derived threshold, the ≥7.0 mg/dL model detected 10.1% (53/527) of female cases in KNHANES; sex-specific thresholds chosen in the internal test set detected 86.3% (455/527) but flagged 47% (13,018/27,706) of women at a PPV of 3.5%, against a female prevalence of 1.9%, so we found no rule that identifies most women with hyperuricemia at a low testing burden. Events at ≥9.0 mg/dL were few (494 in the internal test set and 444 in KNHANES, 45 of them in women). The KNHANES sample is a complete-case subset, and the survey-weighted CIs ignore the sampling design.

Conclusions

Routine dipstick results add a small increment to age, sex, and BMI for identifying serum urate ≥7.0 mg/dL. The models' ranking transferred to a national survey sample, but their absolute risks did not, and at the thresholds tested they mainly selected men. Because both cohorts had blood drawn, their use for deciding who should have serum urate measured in urine-only encounters remains a hypothesis. Before any use, the models would need local recalibration and prospective evaluation in the intended setting.

Acknowledgments

The authors thank the Korea Disease Control and Prevention Agency for providing KNHANES data and the developers of *scikit-learn*, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Shapley additive explanations (SHAP).

Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number RS-2025-02307233). This work was also supported by the Doctoral Student Research Grant Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2025-25400035). The funders had no role in the study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to submit the manuscript for publication.

Data Availability

The KNHANES data are publicly available from the Korea Disease Control and Prevention Agency [18]. The hospital data cannot be shared publicly because of institutional privacy regulations and the Korean Personal Information Protection Act; requests for collaborative access may be directed to the corresponding author and are subject to IRB approval. The code that produced every result in this article, including cohort construction, model training, the KNHANES coding, pinned software versions, and the checksum record made before the models were fitted, is available in the public repository [50]. The fitted model files are not publicly released because of the pending patent application; requests for access for independent validation may be made to the corresponding author. Aggregated results for every analysis are in Multimedia Appendix 2.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (Claude) was used for text generation; proofreading, editing, or language polishing; translation; generating figures or tables; data cleaning; visualization or media generation; code generation or optimization; reference formatting. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript.

Authors' Contributions

ECJ conceptualized the study, developed the methodology, performed the formal analysis, developed the software, created the visualizations, and wrote the original draft. PS, JEK, MHH, and HH contributed to data curation and investigation. KL conceptualized and supervised the study, acquired funding, administered the project, and critically reviewed and edited the manuscript. All authors approved the final version of the manuscript. KL is the guarantor of this work.

XSL•FO
RenderX

Conflicts of Interest

ECJ, KL, PS, and JEK are inventors on a pending Korean patent application (10-2026-0130915, filed July 15, 2026; applicant: Seoul National University Hospital) for an apparatus and method that predict hyperuricemia risk from urinalysis information, which is related to this work. No other competing interests were declared.

Multimedia Appendix 1

TRIPOD+AI checklist and TRIPOD+AI for Abstracts checklist.
[DOCX File , 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Supplementary methods, tables, and figures: study flow and setting (section S1); inputs, input coding, and the Korea National Health and Nutrition Examination Survey (KNHANES) mapping (section S2); missing data, selection, and verification bias (section S3); tuned hyperparameters (section S4); added value of the dipstick and the ensemble versus simpler models (section S5); discrimination and calibration (section S6); clinical utility, including operating points by sex (section S7); transportability over time and population (section S8); subgroups (section S9); nested-threshold coherence and the continuous-urate model (section S10); interpretation of the ensembles (section S11); and the exploratory ≥10.0 mg/dL model (section S12).
[DOCX File , 1128 KB-Multimedia Appendix 2]

References

1. Chen-Xu M, Yokose C, Rai SK, Pillinger MH, Choi HK. Contemporary prevalence of gout and hyperuricemia in the United States and decadal trends: the National Health and Nutrition Examination Survey, 2007-2016. Arthritis Rheumatol. 2019;71(6):991-999. [FREE Full text] [doi: 10.1002/art.40807] [Medline: 30618180]
2. Kuo C, Grainge MJ, Zhang W, Doherty M. Global epidemiology of gout: prevalence, incidence and risk factors. Nat Rev Rheumatol. 2015;11(11):649-662. [doi: 10.1038/nrrheum.2015.91] [Medline: 26150127]
3. Dalbeth N, Choi HK, Joosten LAB, Khanna PP, Matsuo H, Perez-Ruiz F, et al. Gout. Nat Rev Dis Primers. 2019;5(1):69. [doi: 10.1038/s41572-019-0115-y] [Medline: 31558729]
4. Borghi C, Agabiti-Rosei E, Johnson RJ, Kielstein JT, Lurbe E, Mancia G, et al. Hyperuricaemia and gout in cardiovascular, metabolic and kidney disease. Eur J Intern Med. 2020;80:1-11. [doi: 10.1016/j.ejim.2020.07.006] [Medline: 32739239]
5. Johnson RJ, Bakris GL, Borghi C, Chonchol MB, Feldman D, Lanaspa MA, et al. Hyperuricemia, acute and chronic kidney disease, hypertension, and cardiovascular disease: report of a scientific workshop organized by the National Kidney Foundation. Am J Kidney Dis. 2018;71(6):851-865. [FREE Full text] [doi: 10.1053/j.ajkd.2017.12.009] [Medline: 29496260]
6. Sirota JC, McFann K, Targher G, Johnson RJ, Chonchol M, Jalal DI. Elevated serum uric acid levels are associated with non-alcoholic fatty liver disease independently of metabolic syndrome features in the United States: liver ultrasound data from the National Health and Nutrition Examination Survey. Metabolism. 2013;62(3):392-399. [FREE Full text] [doi: 10.1016/j.metabol.2012.08.013] [Medline: 23036645]
7. Cho SK, Chang Y, Kim I, Ryu S. U-shaped association between serum uric acid level and risk of mortality: a cohort study. Arthritis Rheumatol. 2018;70(7):1122-1132. [doi: 10.1002/art.40472] [Medline: 29694719]
8. FitzGerald JD, Dalbeth N, Mikuls T, Brignardello-Petersen R, Guyatt G, Abeles AM, et al. 2020 American College of Rheumatology guideline for the management of gout. Arthritis Care Res (Hoboken). 2020;72(6):744-760. [FREE Full text] [doi: 10.1002/acr.24180] [Medline: 32391934]
9. Singh G, Lingala B, Mithal A. Gout and hyperuricaemia in the USA: prevalence and trends. Rheumatology (Oxford). 2019;58(12):2177-2180. [doi: 10.1093/rheumatology/kez196] [Medline: 31168609]
10. Bardin T, Richette P. Definition of hyperuricemia and gouty conditions. Curr Opin Rheumatol. 2014;26(2):186-191. [doi: 10.1097/BOR.0000000000000028] [Medline: 24419750]
11. Kanbara A, Hakoda M, Seyama I. Urine alkalization facilitates uric acid excretion. Nutr J. 2010;9:45. [FREE Full text] [doi: 10.1186/1475-2891-9-45] [Medline: 20955624]
12. Sakhaee K, Adams-Huet B, Moe OW, Pak CY. Pathophysiologic basis for normouricosuric uric acid nephrolithiasis. Kidney Int. 2002;62(3):971-979. [FREE Full text] [doi: 10.1046/j.1523-1755.2002.00508.x] [Medline: 12164880]
13. Choi HK, Atkinson K, Karlson EW, Curhan G. Obesity, weight change, hypertension, diuretic use, and risk of gout in men: the health professionals follow-up study. Arch Intern Med. 2005;165(7):742-748. [doi: 10.1001/archinte.165.7.742] [Medline: 15824292]
14. Quiñones Galvan A, Natali A, Baldi S, Frascerra S, Sanna G, Ciociaro D, et al. Effect of insulin on uric acid excretion in humans. Am J Physiol. 1995;268(1 Pt 1):E1-E5. [doi: 10.1152/ajpendo.1995.268.1.E1] [Medline: 7840165]
15. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
16. Campion EW, Glynn RJ, DeLabry LO. Asymptomatic hyperuricemia. risks and consequences in the normative aging study. Am J Med. 1987;82(3):421-426. [doi: 10.1016/0002-9343(87)90441-4] [Medline: 3826098]

XSL•FO
RenderX

17. Dalbeth N, Phipps-Green A, Frampton C, Neogi T, Taylor WJ, Merriman TR. Relationship between serum urate concentration and clinically evident incident gout: an individual participant data analysis. Ann Rheum Dis. 2018;77(7):1048-1052. [doi: 10.1136/annrheumdis-2017-212288] [Medline: 29463518]
18. Kweon S, Kim Y, Jang M, Kim Y, Kim K, Choi S, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES). Int J Epidemiol. 2014;43(1):69-77. [FREE Full text] [doi: 10.1093/ije/dyt228] [Medline: 24585853]
19. Oh K, Kim Y, Kweon S, Kim S, Yun S, Park S, et al. Korea National Health and Nutrition Examination Survey, 20th anniversary: accomplishments and future directions. Epidemiol Health. 2021;43:e2021025. [FREE Full text] [doi: 10.4178/epih.e2021025] [Medline: 33872484]
20. Korea National Health and Nutrition Examination Survey: guidelines for the use of raw data, 9th cycle (2022-2024), version 2512 [in Korean]. Korea National Health and Nutrition Examination Survey (KNHANES). URL: https://knhanes.kdca.go.kr/ [accessed 2026-10-07]
21. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-1457. [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
23. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12(10):e1001885. [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
24. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214-226. [FREE Full text] [doi: 10.1002/sim.6787] [Medline: 26553135]
25. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19):4230-4251. [doi: 10.1002/sim.9025] [Medline: 34031906]
26. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146-3154. [FREE Full text]
27. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785-794. [doi: 10.1145/2939672.2939785]
28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31:6639-6649. [FREE Full text]
29. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Proc 8th ACM SIGKDD Int Conf Knowl Discov Data Min. 2002:694-699. [doi: 10.1145/775047.775151]
30. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. [FREE Full text] [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
32. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. [FREE Full text] [doi: 10.1136/bmj-2023-074819] [Medline: 38191193]
33. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61(1):76-86. [doi: 10.1016/j.jclinepi.2007.04.018] [Medline: 18083464]
34. 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):837-845. [Medline: 3203132]
35. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
36. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
37. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, United States. John Wiley & Sons, Inc; 1987. [doi: 10.1002/9780470316696]
38. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39(1):207-215. [Medline: 6871349]
39. de Groot JAH, Bossuyt PMM, Reitsma JB, Rutjes AWS, Dendukuri N, Janssen KJM, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770. [FREE Full text] [doi: 10.1136/bmj.d4770] [Medline: 21810869]

XSL•FO
RenderX

40. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
41. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-4774. [FREE Full text]
42. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279-289. [FREE Full text] [doi: 10.1016/j.jclinepi.2014.06.018] [Medline: 25179855]
43. Lee S, Choe E, Park B. Exploration of machine learning for hyperuricemia prediction models based on basic health checkup tests. J Clin Med. 2019;8(2):172. [FREE Full text] [doi: 10.3390/jcm8020172] [Medline: 30717373]
44. Zeng L, Ma P, Li Z, Liang S, Wu C, Hong C, et al. Multimodal machine learning-based marker enables early detection and prognosis prediction for hyperuricemia. Adv Sci (Weinh). 2024;11(34):e2404047. [doi: 10.1002/advs.202404047] [Medline: 38976552]
45. Ichikawa N, Taniguchi A, Urano W, Nakajima A, Yamanaka H. Comorbidities in patients with gout. Nucleosides Nucleotides Nucleic Acids. 2011;30(12):1045-1050. [doi: 10.1080/15257770.2011.596499] [Medline: 22132955]
46. Jang E, Park Y, Han H, Lee C, Kang E, Lee Y, et al. Machine-learning enhancement of urine dipstick tests for chronic kidney disease detection. J Am Med Inform Assoc. 2023;30(6):1114-1124. [FREE Full text] [doi: 10.1093/jamia/ocad051] [Medline: 37027837]
47. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
48. Lee C, Yoon J, van der Schaar M. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng. 2020;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]
49. Park SP, Yue Y, Li C, Park S. Longitudinal serum uric acid transition patterns and their clinical, genetic, and dietary determinants: prospective prediction of incident hyperuricemia and gout in a Korean population-based cohort. Clin Rheumatol. 2026;45(7):4509-4526. [doi: 10.1007/s10067-026-08182-z] [Medline: 42240788]
50. Source code repository for the urine dipstick hyperuricemia screening models. GitHub. URL: https://github.com/eunchanj/Urine-Dipstick-Hyperuricemia-Screening [accessed 2026-10-17]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CatBoost: categorical boosting
IRB: institutional review board
KDCA: Korea Disease Control and Prevention Agency
KNHANES: Korea National Health and Nutrition Examination Survey
LightGBM: Light Gradient Boosting Machine
O:E: observed-to-expected
OR: odds ratio
PPV: positive predictive value
RECORD: Reporting of Studies Conducted Using Observational Routinely Collected Health Data
SHAP: Shapley additive explanations
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis extension for Artificial Intelligence
XGBoost: Extreme Gradient Boosting

Edited by I Steenstra; submitted 20.Jul.2026; peer-reviewed by N Itzhak, J Yoon; comments to author 28.Sep.2026; revised version received 24.Oct.2026; accepted 26.Oct.2026; published 20.Nov.2026

Please cite as:
Jang EC, Song P, Kim JE, Han MH, Han H, Lee K
Machine Learning Models Combining Routine Urine Dipstick Results With Age, Sex, and BMI to Triage Serum Urate Testing: Development and External Validation Study
JMIR Med Inform 2026;14:e107515
URL: https://medinform.jmir.org/2026/1/e107515
doi: 10.2196/107515
PMID:

XSL•FO
RenderX

©Eun Chan Jang, Pado Song, Joung Eun Kim, Myung Hwa Han, Hwigyeong Han, Kiheon Lee. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX