Machine Learning Models to Predict 24 Hour Urinary Abnormalities for Kidney Stone Disease

Nicholas L. Kavoussi, Chase Floyd, Abin Abraham, Wilson Sui, Cosmin Bejan, John A. Capra, and Ryan Hsi

OBJECTIVE To help guide empiric therapy for kidney stone disease, we sought to demonstrate the feasibility of predicting 24-hour urine abnormalities using machine learning methods.

METHODS We trained a machine learning model (XGBoost [XG]) to predict 24-hour urine abnormalities from electronic health record-derived data (n = 1314). The machine learning model was compared to a logistic regression model [LR]. Additionally, an ensemble (EN) model combining both XG and LR models was evaluated as well. Models predicted binary 24-hour urine values for volume, sodium, oxalate, calcium, uric acid, and citrate; as well as a multiclass prediction of pH. We evaluated performance using area under the receiver operating curve (AUC-ROC) and identified predictors for each model.

RESULTS The XG model was able to discriminate 24-hour urine abnormalities with fair performance, comparable to LR. The XG model most accurately predicted abnormalities of urine volume (accuracy = 98%, AUC-ROC = 0.59), uric acid (69%, 0.73) and elevated urine sodium (71%, 0.79). The LR model outperformed the XG model alone in prediction of abnormalities of urinary pH (AUC-ROC of 0.66 vs 0.57) and citrate (0.69 vs 0.64). The EN model most accurately predicted abnormalities of oxalate (accuracy = 65%, ROC-AUC = 0.70) and citrate (65%, 0.69) with overall similar predictive performance to either XG or LR alone. Body mass index, age, and gender were the three most important features for training the models for all outcomes.

CONCLUSION Urine chemistry prediction for kidney stone disease appears to be feasible with machine learning methods. Further optimization of the performance could facilitate dietary or pharmacologic prevention. UROLOGY 169: 52−57, 2022. © 2022 Elsevier Inc.

Kidney stones are common and after an index stone event, recurrence rates can be as high as 50%.[1,2] Clinical guidelines recommend metabolic testing with one or two 24-hour (24H) urine collections to guide selective pharmacologic and dietary interventions to mitigate repeat stone events.[3] In practice, however, 24H urine collection rates are low overall.[4] In addition, testing may not be available in certain health care settings or covered by insurance, while others do not have access to testing. Accurate methods for prediction of 24H urine analyte abnormalities by using demographic and clinical electronic health record-derived (EHR) data have the potential to enable identification of patients for whom 24H urine testing would most likely reveal an abnormality that could be targeted by empiric dietary or pharmacotherapy.

Prior studies have demonstrated limited accuracy of logistic regression (LR, 64%) in the prediction of 24H urine parameters using EHR-derived data.[5] Machine learning methods may be preferable for the analysis of EHR data and improve the prediction of 24H urine parameters. Specifically, machine learning models, such as boosted decision trees, build mathematical algorithms from raw, labeled training data to classify predictor significance. Training of these algorithms is not predicated on predictor significance, making machine learning techniques robust in the analysis of non-linear data, such as EHR data. We previously demonstrated the utility of machine learning techniques in prediction of stone composition from clinical parameters.[6] However, machine learning methods have not been created for the prediction of 24H urine abnormalities from EHR-derived data.

Funding Support: *The project described was supported by CTSA award No. UL1TR000445 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.*

From the Department of Urology, Vanderbilt University Medical Center, Nashville, TN; the University of South Carolina School of Medicine, Columbia, SC; the Department of Biological Sciences, Vanderbilt Genetics Institute, and Center for Structural Biology, Vanderbilt University, Nashville, TN; the Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; the Bakar Computational Health Sciences Institute at the University of California, San Francisco, CA; and the Department of Epidemiology and Biostatistics at the University of California, San Francisco, CA

Address correspondence to: Nicholas L Kavoussi, MD, Department of Urology, Vanderbilt University Medical Center, 1211 Medical Center Drive, Nashville, Tennessee, 37212. E-mail: Nicholas.l.kavoussi@vumc.org

Submitted: April 12, 2022, accepted (with revisions): July 4, 2022

https://doi.org/10.1016/j.urology.2022.07.008
0090-4295

Within this context, we sought to develop machine learning models for predicting 24H urine abnormalities from EHR- derived data using a single-institution cohort of patients with kidney stone disease. To accomplish this, we trained and compared both a boosted decision tree (XGBoost, XG) and a logistic regression (LR) machine learning models from EHR-derived clinical and demographic data to predict 24H urine abnormalities. We additionally sought to identify which clinical and demographic predictors most significantly predicted 24H urine abnormalities in each model.

MATERIALS AND METHODS

Patient Cohort

After local institutional review board approval, we performed a retrospective review of all adult patients with kidney stone disease who completed 24H urine evaluation at our institution between 2009-2019 (N = 1314). We extracted demographic and clinical data using a semi-automated data extraction tool from our cohort using our an institutionally maintained database of the entire electronic health record.[7–9] We extracted demographic and clinical data from the EHR using a semiautomated data extraction tool.[10,11] Demographic information recorded included age at time of 24H urine testing, gender, body mass index (BMI) and race. We extracted clinical predictors associated with kidney stone risk based on the International Classification of Disease coding (See Appendix Table 1). Stone composition analysis was determined from an external laboratory using infrared spectroscopy (Beck Laboratories, Greenwood, IN). If stone composition was mixed, the stones were categorized by highest percentage composition. We also identified whether patients had been prescribed an alkalinizing agent, allopurinol or a thiazide diuretic (see Appendix Table 2).

24H Urine Parameters

All 24H urine testing was conducted by an external laboratory (Litholink Corporation, Chicago, IL), and only adequate collections based on sex-specific creatinine per kilogram (Cr24/kg) measurements were included in this study. If patients had multiple 24H urine studies, we identified the study temporally closest to a stone event (ie stone surgery or spontaneous passage of stone). We extracted the following individual urine parameters from the 24H urine study: urine volume (Vol24) calcium (Ca24), oxalate (Ox24), citrate (Cit24), UA (UA24), sodium (Na24), and urine pH.[4]

Models for Predicting 24H Urine Abnormalities

We evaluated whether a gradient boosted decision tree (XGBoost version 0.81, XG) could predict 24H urine abnormalities. XGBoost leverages many decision trees for prediction and penalizes incorrect predictions from previous decision trees.[12] Boosted decision trees are particularly robust for the non-linear correlation of predictors, such as EHR-derived data.

We trained our XG model using the EHR-derived predictors described above. Race and gender were categorically encoded. Clinical predictors and medication exposure were encoded as binary variables. All other predictors were treated as continuous variables.

Then, for the outcomes of interest, we classified values for volume, sodium, oxalate, calcium, and uric acid binarily (ie normal vs high), as well as citrate (ie normal vs low). We categorized pH values via multiclass categorization (ie low, normal, high). The reference ranges for laboratory values were derived from prior studies evaluating 24H urine chemistries as used by the specialized laboratory used for analysis (Litholink Corporation, Chicago, IL) (Appendix Table 3).[13]

We randomly divided the data with equal proportions of each urine abnormality into a training (80%) and validation cohort (20%). Standardized hyperparameters were optimized using Bayisan techniques.[14]

We trained logistic regression (LR) models using the same cohort for the XG models but did not perform any hyperparameter tuning. We evaluated LR model performance using the respective validation cohort. As XG and LR may each individually rely on few predictors, we evaluated whether or not combining XG and LR in an ensemble model (EN) would improve performance.[13] Specifically, we trained EN using predictions from the initial XG and LR models as inputs into a meta-model (also logistic regression) to predict 24H urine abnormalities (Fig. 1). Then, we similarly evaluated the EN models using the validation cohort.

Evaluation Metrics

The primary outcomes were accuracy and area under the receiver operating curve (AUC-ROC) for the prediction of 24H urine abnormalities for each model. Secondary outcomes included the significance of each EHR-derived variable used to train the models via Shapley Additive Explanation (SHAP, v0.35) score.[14] SHAP scores represent the relative contribution of each predictor used for classification based on the log-odds units of change in prediction. All analysis was conducted with Python v3.8.[15]

RESULTS

The EHR-derived predictors used for model training are presented in Table 1. Of the 1314 patients included for analysis, the patients were primarily white (91%) with the most common comorbidities being hypertension (54%), gastroesophageal reflux disease (GERD, 36%), and hyperlipidemia (28%). Only a minority of patients were on any medical therapy for kidney stone disease with an alkalinizing agent (9%), a thiazide diuretic (7%) or allopurinol (4%). Predominant stone compositions included calcium oxalate (67%), hydroxyapatite (18%), carbonate apatite (2%), uric acid (8%), and other (ie struvite or cystine stones, 5%). The most common abnormality seen on 24H urine testing was high urine sodium (61%), followed by hypocitraturia (45%), low urine pH (44%), hypercalciuria (41%), hyperoxaluria (37%), high urine pH (31%), and hyperuricosuria (23%) (Appendix Table 3).

We evaluated the XG and LR models' ability to classify each identified 24H urine parameter as normal or abnormal, and then combined them in an EN model (Fig. 1). The XG model was able to discriminate 24-hour urine abnormalities with fair performance that was comparable to LR (Fig. 2, Appendix Fig. 1). The XG model most accurately predicted abnormalities of urine volume (accuracy = 98%, AUC-ROC = 0.59), uric acid (69%, 0.73) and elevated urine sodium (71%, 0.79). The LR model outperformed the XG model alone in prediction of abnormalities of urinary pH (AUC-ROC of 0.66 vs 0.57) and citrate (0.69 vs0.64). The EN model most accurately predicted abnormalities of oxalate (accuracy = 65%, ROC-AUC = 0.70) and citrate (65%, 0.69)

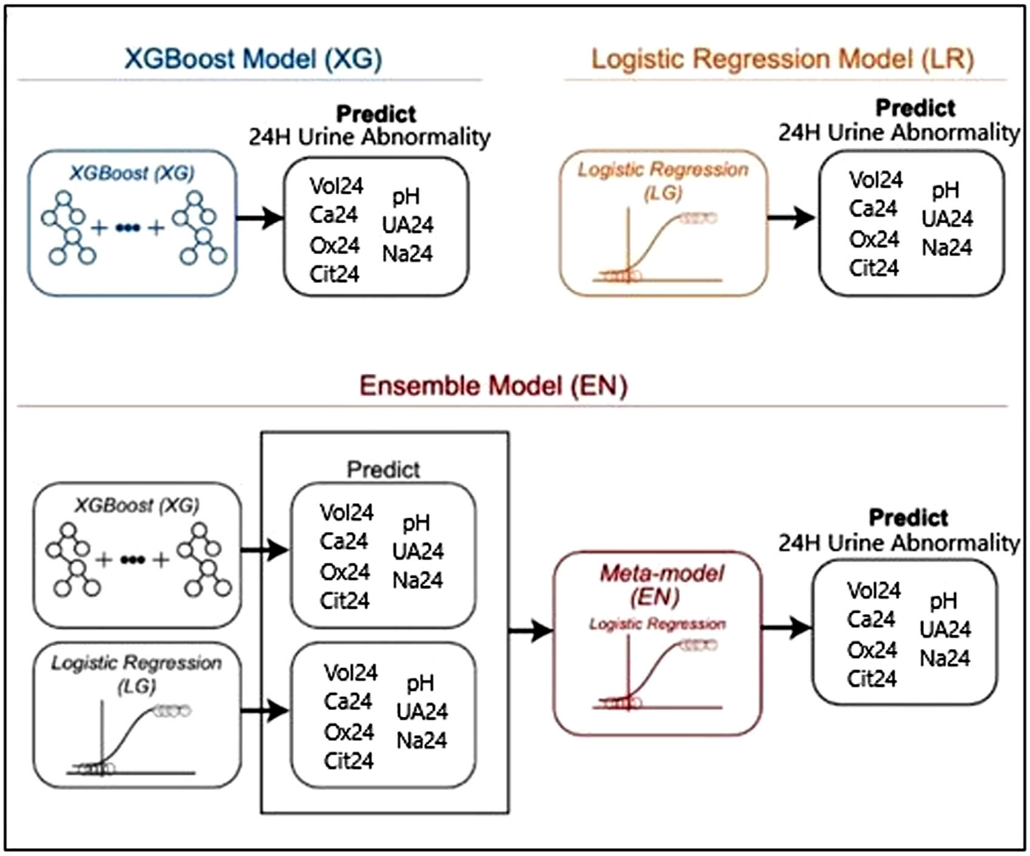

Figure 1. Workflow for prediction using each model. Each model underwent training using EHR-derived data. (Color version available online.)

with overall similar predictive performance to either XG or LR alone.

For each 24H urine abnormality, different sets of predictors were prioritized for the final prediction by the XG model (Table 2). Overall, BMI was the most significant predictor for each 24H urine abnormality. Other top predictors included age, male gender, and calcium oxalate stone composition. Race was not deemed to be an important predictor of 24H urine abnormality by the model.

DISCUSSION

We found that machine learning models can feasibly predict metabolic urine chemistry abnormalities of kidney stone patients using demographic, clinical, and stone composition predictors. Specifically, the XG model outperformed the LR model for prediction of abnormalities of urine volume, uric acid and sodium concentrations. However, the LR model outperformed the XG model in prediction of abnormalities of urinary citrate and pH. Combination of the models in the EN model added some improvement in the performance above the XG model. Though machine learning models can analyze complex, non-linear patterns in datasets, the weaker performance of the XG model may be indicative of the need for larger samples sizes for model training. In addition, the accuracy measurement depends on the threshold chosen, whereas the AUC-ROC is invariant to thresholds. Therefore, depending on the clinical question, different models can be chosen to provide the best performing model. For example, future models could test the potential appropriateness of thiazides versus potassium citrate based on demographic and clinical history factors. Finally, we also found that the demographic and clinical predictors prioritized by the XG model for training reflect known associations with stone disease, such as BMI. Together, these findings support the potential of EHR derived tools for predicting 24H urine abnormalities in kidney stone patients.

EHR-derived tools that associate predictors with 24H urine abnormalities could facilitate earlier, selective dietary and medical interventions for stone disease. Though recent work has suggested that selective medical therapy may have similar impact on clinical stone recurrence as empiric medical therapy, selective medical therapy with metabolic urine testing can benefit patients with risk factors for stone recurrence.[15,16] Metabolic urine testing for stone prevention with 24H urine analysis can be limited both by patient collection and manual clinical interpretation.[16] Additionally, the cost of the test and its utility in individualized stone prevention have led to underutilization of metabolic urine testing despite recommendations

Table 1. Patient characteristics serving as inputs for model training

Demographics	N = 1314 (%)
Age (years ± SD)	51±15
Gender, male	697 (53)
Gender, female	617 (47)
BMI (mean ± SD)	30±8
Race	
White	1192 (91)
African American	56 (4)
Asian	20 (2)
Other	40 (3)
Past Medical History	
Bowel Disease, N(%)	119 (9)
Hyperlipidemia, N(%)	368 (28)
Hypertension, N(%)	707 (54)
Gout, N(%)	57 (4)
Diabetes, N(%)	292 (22)
Chronic Kidney Disease, N(%)	105(8)
Cystinuria, N(%)	3 (0.2)
Coronary Artery Disease, N(%)	130 (10)
Cerebrovascular Accident, N(%)	36 (3)
Gastroesophageal Reflux Disease, N(%)	469 (36)
Osteoporosis, immobility or hyperparathyroidism, N(%)	72 (5)
Meds	
Alkalinizing agent, N(%)	112 (9)
Thiazide, N(%)	89 (7)
Allopurinol, N(%)	52 (4)
Predominant Stone Composition*	
Calcium Oxalate, N(%)	880 (67)
Hydroxyapatite, N(%)	241 (18)
Carbonate Apatite, N(%)	20(2)
Uric Acid, N(%)	100 (8)
Other, N(%)	60 (5)

* Stone composition only available for 1301 patients in cohort.

by the American Urologic Association.[3,13,17,18] Our results suggest the feasibility of a machine learning based prediction tools to identify interventions to target the urinary parameters most likely to be abnormal.

Prior studies have investigated the prediction of 24H urine abnormalities from EHR-derived data. Otto et al associated demographic and clinical information with 24H urine stone risk results.[5] Specifically, in calcium oxalate stone formers the group found that age, gender, and BMI all impact calcium, oxalate, citrate, and pH using logistic regression models. Our boosted machine learning model similarly prioritized these variables for 24H urine abnormality prediction. However, our inclusion of all stone compositions (ie not only calcium oxalate) likely explains differences in prediction of specific 24H urine analytes seen in our study. Moreover, it is likely that more robust data sets for training of the machine learning models will improve prediction.

The predictors prioritized by our models support the current understanding of stone pathophysiology. Age, gender, and BMI all have known physiologic impacts on urine analytes. For example, insulin resistance in the proximal tubule found in obese patients can lead to decreased ammonia excretion and the lithogenic acidification of urine.[19] We similarly found BMI to be the top prioritized predictor of 24H urine pH prediction. Likewise, known age related changes to GFR may associate with type 4 renal tubular acidosis, favoring more acidic urine.[18] Finally, the mediation of calcium excretion by estrogen has been associated with differences in stone compositions between men and women.[20,21] Future efforts to optimize machine learning models as clinic decision making tools could prioritize these clinical and demographic EHR-derived variables for improved prediction. Additionally, further studies assessing the utility of machine learning models to guide empiric medical management compared to selective medical therapy based on 24H urine results are pending.

There are limitations to our study. First, the study's retrospective data collection may not account for variables missing from the EHR, and there may be important predictors that were not included in the models. Additionally, it is possible that some patients had modified dietary, pharmacologic, and lifestyle factors that would have impacted their urine metabolites at the time of the 24H study. Notably, the inclusion of patients on thiazides and alkali citrate, which are known to improve urine chemistries, may confound our results. However, these patients were few (16%), and these medications for reasons other than stone disease (eg 90% of patients taking thiazides

Table 2. Top XGBoost predictors of 24H urine abnormalities

		24H Urine Parameter						
		pH	UA24	Na24	Cit24	Ca24	Ox24	Vol24
Top XG Boost Predictors*	BMI	0.42	0.31	0.29	0.15	0.04	0.004	0.005
	Age	0.29	0.16	0.05	0.13	0.03	0.003	0.004
	Male gender	0.02	0.06	0.21	0.01	0.001	0.001	0.001
	Ca Oxalate stone	0.16	0.01	0.008	0.11	0.003	0.0004	0.001
	Female gender	0.05	0.11	0.08	0.04	0.003	0.007	0.001
	Hypertension	0.04	0.02	0.03	0.08	0	0	0.0004
	GERD	0.07	0.04	0.02	0.009	0.001	0.0002	0.001
	Diabetes mellitus	0.06	0.01	0.01	0.01	0.001	0.001	0.0002
	Hydroxyapatite stone	0.07	0.01	0	0.002	0.001	0	0.0003
	Thiazide medication	0.02	0.004	0	0	0	0	0.001

* Predictors are sorted from most to least significant by SHAP score, representing the relative contribution of each predictor used for classification via the change in log-odds. A score of 0 suggests the predictor has no influence on abnormal urinary parameter prediction. A non-zero score suggests association with predicting an abnormal urinary parameter.

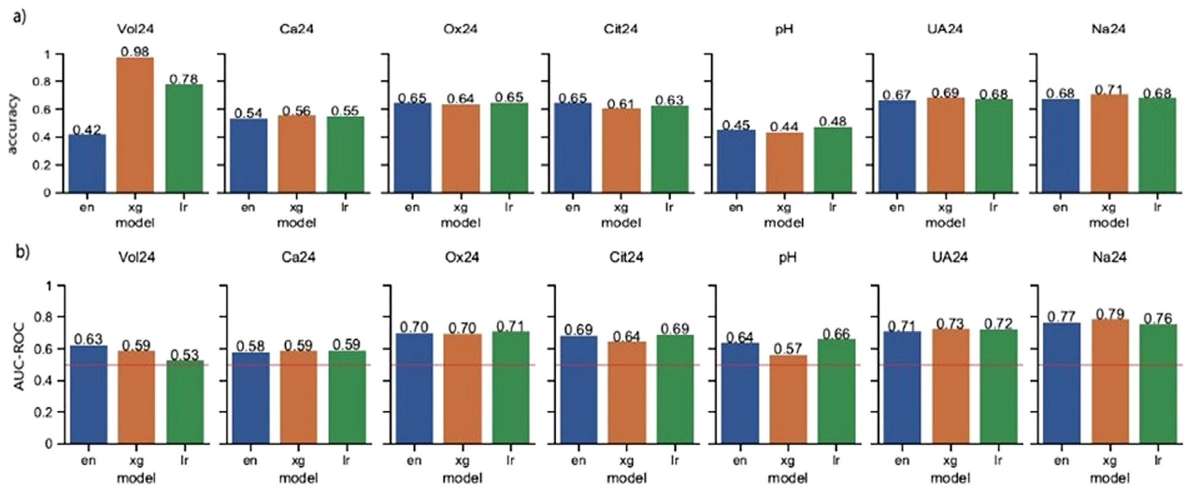

Figure 2. Performance of each model for predicting 24H urine abnormalities. A) accuracy. B) AUC-ROC. EN=Ensemble Model, XG=XGBoost, LR=Logistic Regression. (Color version available online.)

were on the medications for hypertension). Furthermore, medication status was not highly prioritized by the model as prediction variables for any urinary parameter. We chose to evaluate a single test that was temporally closest to a stone event to minimize these effects, but this could also limit model training. Moreover, there are likely other clinical characteristics that could be included in the models that was inconsistently reported in the EHR and we were unable to use (ie age of first stone, number of stone episodes etc.) Furthermore, the reference values for urine parameters are somewhat arbitrary and are limited in directing therapy for stone prevention as even patients with normal urinary parameters may still form stones. Due to this, a binary stratification of "abnormal vs normal" was used, which may limit granularity and prediction of the models. The findings of this study may differ from data derived from other populations. Despite these limitations, this study demonstrates the feasibility of EHR-derived prediction tools to detect metabolic abnormalities. Further optimization of the models, as well as external validation, can help with clinical decision making for earlier, targeted stone prevention therapy.

CONCLUSION

We have developed machine learning models for the prediction of 24H urine abnormalities using EHR-derived data. Predictors prioritized by our models support the current understanding of kidney stone pathophysiology. Further studies aimed at model optimization and validation could lead to the creation of clinical tools to facilitate decision -making for medical stone management.

DECLARATION OF COMPETING INTEREST

None.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.urology.2022.07.008.

References

1. Fink HA, Wilt TJ, Eidman KE, Garimella PS, MacDonald R. Recurrent nephrolithiasis in adults: comparative effectiveness of preventive medical strategies. *Agency Healthc Res Qual*. 2012;61:232.
2. Uribarri J, Oh MS, Carroll HJ. The first kidney stone. *Ann Intern Med*. 1989;111:1006–1009. https://doi.org/10.7326/0003-4819-111-12-1006.
3. Pearle MS, Goldfarb DS, Assimos DG, et al. Medical management of kidney stones: AUA guideline. *J Urol*. 2014;192:316–324. https://doi.org/10.1016/j.juro.2014.05.006.
4. Milose JC, Kaufman SR, Hollenbeck BK, Wolf JS, Hollingsworth JM. Prevalence of 24-hour urine collection in high risk stone formers. *J Urol*. 2014;191:376–380. https://doi.org/10.1016/j.juro.2013.08.080.
5. Otto BJ, Bozorgmehri S, Kuo J, Canales M, Bird VG, Canales B. Age, body mass index, and gender predict 24-hour urine parameters in recurrent idiopathic calcium oxalate stone formers. *J Endourol*. 2017;31:1335–1341. https://doi.org/10.1089/end.2017.0352.
6. Abraham A, Kavoussi N, Sui W, Bejan C, Capra JA, Hsi R. Machine learning prediction of kidney stone composition using electronic health record-derived features. *J Endourol*. 2021. Published online July 27. https://doi.org/10.1089/end.2021.0211.
7. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform*. 2014;52:28–35. https://doi.org/10.1016/j.jbi.2014.02.003.
8. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95: 103208. https://doi.org/10.1016/j.jbi.2019.103208.
9. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–381. https://doi.org/10.1016/j.jbi.2008.08.010.
10. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–381.

11. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95: 103208.
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. 2016. https://doi.org/10.1145/2939672.2939785.
13. Curhan GC, Willett WC, Speizer FE, Stampfer MJ. Twenty-four-hour urine chemistries and the risk of kidney stones among women and men. *Kidney Int*. 2001;59:2290–2298. https://doi.org/10.1046/j.1523-1755.2001.00746.x.
14. Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. *Presented at the 30th International Conference on Machine Learning (ICML 2013), Atlanta, Gerorgia, June 16 – 21, 2013. In JMLR Workshop and Conference Proceedings*. 2013;28:115–123.
15. Hsi RS, Yan PL, Goldfarb DS, et al. Comparison of selective versus empiric pharmacologic preventative therapy with kidney stone recurrence. *Urology*. 2021;149:81–88. https://doi.org/10.1016/j.urology.2020.11.054.
16. Hsi RS, Yan PL, Crivelli JJ, Goldfarb DS, Shahinian V, Hollingsworth JM. Comparison of selective vs empiric pharmacologic preventive therapy of kidney stone recurrence with high-risk features. *Urology*. 2022. Published online February 17S0090-4295(22)00140-6. https://doi.org/10.1016/j.urology.2021.12.037.
17. Maalouf NM, Sakhaee K, Parks JH, Coe FL, Adams-Huet B, Pak CYC. Association of urinary pH with body weight in nephrolithiasis. *Kidney Int*. 2004;65:1422–1425. https://doi.org/10.1111/j.1523-1755.2004.00522.x.
18. Lieske JC, Rule AD, Krambeck AE, et al. Stone composition as a function of age and sex. *Clin J Am Soc Nephrol CJASN*. 2014;9:2141–2146. https://doi.org/10.2215/CJN.05660614.