Title: Kidney Stone Decision Engine accurately predicts kidney stone removal and treatment complications for shock wave lithotripsy and laser ureterorenoscopy patients

Authors: Peter A Noble[1*], Blake D. Hamilton[2], Glenn Gerber[3].

Affiliations:

[1]University of Alabama Birmingham, Department of Microbiology, 845 19th Street South Birmingham, AL 35294.

[2]School of Medicine, University of Utah, 50 Medical Dr. N, Salt Lake City, USA 84132

[3]University of Chicago Medical Center, 5841 S. Maryland Ave, Chicago, IL 60637

*Correspondence to: Peter A Noble: panoble2017@gmail.com, Phone: 206-409-6664.

EMAILS:
Peter A Noble: panoble2017@gmail.com; panoble@uab.edu; peteranoble@buddyengineer.com
Blake D. Hamilton: blake.hamilton@hsc.utah.edu
Glenn Gerber: glgerber528@gmail.com

DRAFT VERSION DATE: November 15, 2023

TARGET JOURNAL:

Attachments: Supplementary Information

ABSTRACT (261 words):
Kidney stones form when mineral salts crystallize in the urinary tract. While most stones exit the body in the urine stream, some can block the ureteropelvic junction or ureters, leading to severe lower back pain, blood in the urine, vomiting, and painful urination. Imaging technologies, such as X-rays or ureterorenoscopy (URS), are typically used to detect kidney stones. Subsequently, these stones are fragmented into smaller pieces using shock wave lithotripsy (SWL) or laser URS. Both treatments yield subtly different patient outcomes. To predict successful stone removal and complication outcomes, Artificial Neural Network models were trained on 15,126 SWL and 2,116 URS patient records. These records include patient metrics like Body Mass Index and age, as well as treatment outcomes obtained using various medical instruments and healthcare professionals. Due to the low number of outcome failures in the data (e.g., treatment complications), Nearest Neighbor and Synthetic Minority Oversampling Technique (SMOTE) models were implemented to improve prediction accuracies. To reduce noise in the predictions, ensemble modeling was employed. The average prediction accuracies based on Confusion Matrices for SWL stone removal and treatment complications were 84.8% and 95.0%, respectively, while those for URS were 89.0% and 92.2%, respectively. The average prediction accuracies for SWL based on Area-Under-the-Curve were 74.7% and 62.9%, respectively, while those for URS were 77.2% and 78.9%, respectively. Taken together, the approach yielded moderate to high accurate predictions, regardless of treatment or outcome. The models were incorporated into a Kidney Stone Decision Engine web application (http://peteranoble.com/webapps.html) that suggests the best interventions to healthcare providers based on individual patient metrics.

INTRODUCTION

The incidence and prevalence of kidney stones in people is increasing globally presumably due to dietary practices and global warming [1]. In the United States, about 11% of the population will have kidney stones in their lifetime [2]. The increasing incidence of kidney stones presents a dilemma to healthcare professionals because the 'optimal' intervention to remove the stones varies by approach [3], patient health, age, preference, and body size [4-8], stone size and composition [9], and stone location [10].

Two interventions most often used to remove/fragment stones, include: shock wave lithotripsy (SWL) and laser ureterorenoscopy (URS). SWL uses high-energy shock waves to fragment stones into small particles that eventually pass out of the body in urine [11]. This intervention is a less invasive than URS but not as effective in terms of attaining stone-free status – that is, patients might require additional treatments [12]. A laser attached to the URS is used to fragment stones, which are subsequently either transported out of the body in the urine stream or removed during the procedure [13]. Two drawbacks of URS are: higher incidence of treatment complications and more costly, sometimes requiring longer hospital stays than patients treated by SWL [14,15]. A survey of intervention decisions suggests most patients prefer SWL to URS [16] and a recent Evidence Review by NIH states only 'small benefits of URS over SWL' -- yet clinical and cost effectiveness favor SWL [17]. Selecting the 'optimal'

intervention for patients is therefore not straightforward; an approach that helps healthcare professionals with these decisions is highly desired.

The objective of this study was to build a Kidney Stone Decision Engine (KSDE) based on mining a database containing information on previous interventions (SWL and URS). The databases include information on patient metrics (such as age and Body Mass Index (BMI), stone removal successes/failures, and evidence of treatment complications. We determined the prediction probabilities for various treatment outcomes based on these metrics and the uncertainty of the predictions by repeated independent statistical analyses. Artificial neural network (ANN) models were used to find patterns in the 17242 patient records. The equations of forty models were extracted and incorporated into a KSDE application that healthcare professionals can use in patient counseling to predict SWL or URS outcomes based on patient metrics.

MATERIALS AND METHODS

Electronic Medical Data An existing anonymized database (Kidney Stone Registry) was used. The database consisted of 80,000+ patients who had undergone SWL or URS treatments at multiple sites throughout the United States. We selected 20,000 patient records between February 19th 2018 and August 31st, 2021. We then excluded records with missing or erroneous data to end up with 17242 patient records. Individual patient consent was not required as no patient identifiable records were used in the study.

A variety of SWL and URS instruments were used to treat patients. Specifically, SWL was performed using the Dornier Compact Delta II (DCD2), Dornier Compact Delta III (DCD3), Dornier Compact Sigma (DCS) (Weßling, Germany), Storz F2 (SF2), or Storz SLX-T (SSLXT) instruments by experienced physicians. Laser URS was performed using Dornier Medilas H20 DMH20, Dornier Medilas H30 (DMH30), Dornier Medilas H35 (DMH35), Lumenis Versapulse 100 watt (LV100) (San Jose, CA), Lumenis Versapulse 20 watt (LV20), or Odyssey Convergent 30 watt (OC30) (Alameda, CA) instruments by experienced physicians.

Coding of variables in the data sets.
SWL data. Label, coding, units: Anticoagulants used prior to treatment (True: 0, False: 1), DCD2 (True: 1, False: 0), DCD3 (True: 1, False: 0), DCS (True: 1, False: 0), SF2 (True: 1, False: 0), SSLXT (True: 1, False: 0), Stone location in ureters (True: 1, False: 0), Stone location in kidney (True: 1, False: 0), Stone not specifically located in the kidney or ureters (0ther location) (True: 1, False: 0), Sex (Male: 1, Female: 0), Body Mass Index (BMI, kg/m^2), Age of the patient at time of the procedure (years), Stone width (mm), Stone length (mm), Stone side (Left: 0, Right,1), Other medical conditions (e.g., Diabetes or other without diabetes, True: 1, False: 0).

URS data. Label, coding, units: Sex (Male: 1, Female: 0), DMH20 (True: 1, False: 0), DMH30 (True: 1, False: 0), DMH35 (True: 1, False: 0), LV100 (True: 1, False: 0), LV20 (True: 1, False: 0), OC30 (True: 1, False: 0), Age of the patient at time of the procedure (years), BMI (kg/m^2).

Target outcomes. Two definitions of stone removal outcomes were used: (i) 'stone free' or stone fragments < 4 mm were assigned a value of '0', and (ii) stone fragments > 4mm or 'no change in stone size' were assigned a value of '1'. These outcomes were determined by a physician's review of the follow-up X-ray images and confirmed with patient records indicating no further treatment was required. There were two definitions of 'treatment complications': (i) a patient with 'no complication' was assigned a value of '0', and (ii) a patient with a treatment complication was assigned a value of '1'. Typical treatment complications included pain, fever, urinary tract infection, hematoma, post-operational bleeding, "*steinstrasse*", prolonged dysuria, ureteral perforation, burning, hydronephrosis, acute kidney injury, tachycardia, prolonged gross hematuria, and obstructing fragments.

Standardization of the data. Prior to building the ANN models, continuous variables were standardized by their corresponding minimum (min) and maximum (max) with the formula:

standardized variable = (raw variable – variable min)/ (variable max – variable min)

ANN modeling. The data sets were randomly split into 70% training, 15% testing, and 15% validation. The architecture of the ANN models consisted of an input, a hidden, and an output layer. The number of neurons in input layer was dependent on the number of input variables. The optimal number of neurons in the hidden layer was empirically determined by selecting a range of numbers (e.g., the square root of the number of inputs to the actual number of inputs) and assessing model accuracy using a Confusion Matrix (i.e., (True positives + True negatives)/(False Positives + False Negatives + True Positives + True Negatives). The output layer consisted of a single neuron, the target variable (i.e., stone removal success or treatment complication). In some cases, the model accuracy was assessed by including all data (i.e., training, testing and validation data sets) into the Confusion Matrix, while in others, only the combined testing/validation data sets were used, as specified in the Results below. The Neuroet package [18] settings used for training were as follows: scaling method, standard linear function (0, 1); transfer function for input and output neurons, Log-Sigmoid; training method, Levenberg-Marquardt. Training was automatically stopped when the global error between outputs and targets was minimized after several iterations. Weights and biases were retained to build the final equations in MS Excel and C++ programs.

Balanced and SMOTED data. Preliminary studies showed that the ANN models had difficulties in learning the decision boundaries due to severe imbalances of the data (results not shown). For example, more patient records had successful stone removals than unsuccessful ones and even fewer patient records involved treatment complications. To address this issue, two data augmentation approaches were used: (i) balancing the training data set with equal number of records for each group (i.e., equal number of successes and failures), and (ii) increasing the number of records in the minority class by synthesizing data using Synthetic Minority Oversampling Technique (SMOTE) [19].

The balanced data set approach involved randomly selecting *x* number of records from the majority class to make them equal in number to those in the minority class. The SMOTE

approach involved: (i) splitting the standardized data set into 70% training and 30% testing/validation, and retaining the training data, (ii) using a Nearest Neighbor model (*k*=3 to 5) to select data points in the minority class and drawing vectors between neighboring points; and (iii) randomly generating synthetic data along the vectors until the number of records in the minority equal the number of records in the majority.

The training data from the approaches were then used to build the ANN models. The weights and biases of each ANN model were incorporated into equations in C++.

Suggested Intervention

The intervention was calculated by scoring the predicted averages and standard deviations for successful stone removal and treatment complications. The scoring system was as follows: an average prediction <0.5 was scored as 0; a standard deviation <0.25 was scored as 0; an average prediction >=0.5 was scored as 1; and a standard deviation that was >=0.25 was scored as 1. The scores for SWL stone removal and treatment complications were summed, as were the scores for URS stone removal and treatment complications. If the sum of SWL was greater than the sum of URS, then the suggested intervention was "URS". If the sum of URS was greater than the sum of SWL, then the suggested intervention was "SWL". If the sum of both SWL and URS were 0, then the suggested intervention was 'SWL or URS". If the sum of SWL and URS was greater or equal to 5 then the suggested intervention was "Uncertain".

Statistical and data analyses

Averages, standard deviations, and one- or two- tailed Student T-tests were implemented in Excel spreadsheets. One-tailed T-tests were used when direction of the test was relevant and two-tailed T-tests when the direction of the test was unknown. The data was SMOTED using Jupyter notebooks running Python libraries. All ANN models were built and tested using the bench marked Neuroet package downloaded from http://peteranoble.com/software.html. Library (pROC) in the R-program 4.1.2 (2021-11-01) was used to calculate Area-under-the Curve (AUC).

RESULTS

Descriptive statistics. The Storz SLX-T instrument was more represented (55.9%) in the SWL data set than the Storz F2 (30.7%) and Dornier instruments (13.4%) (Table 1). Also, more stones were in the kidney (i.e., Lower, Mid, Upper Calyx, Pelvis, and Ureterovesical Junction; 77.8%) than the ureters (Lower, Mid, Upper Ureters and Ureteral Pelvic Junction; 21.6%) or other locations (Bladder, Calcified Stent, Pancreas and Staghorn; <1.0%). Slightly more than half of the patients were overweight healthy males with an average age of 57 years and kidney stones of 8 to 9 mm in diameter. Treatment complications were relatively low (<5 %) and most kidney stones (84.4%) were successfully removed by SWL.

The Lumenis Versapulse (100 watt and 20 watt) instruments were more represented (63.3%) in the URS data set than other instruments (36.7%) (Table 2). The composition of the patients was similar to those treated by SWL (Table 1) with slightly more than half being overweight males

with an average age of 57 years. Treatment complications were relatively low (<5 %) and most (92.8%) kidney stones were successfully removed by URS.

ANN model architecture. Tests of ANN model architectures for the balanced and SMOTED data sets revealed 16 hidden neurons were optimal for SWL models and 5 to 7 hidden neurons were optimal for the URS models (results not shown).

Balanced data sets. Models trained with the balanced data set yielded reasonable prediction accuracies ranging from 72.4 to 92.8% for Confusion Matrices and 77.3 to 95.9% for AUC values (Table 3, Tables S1-S8, Figures S1-S2). However, when the same models were tested on the entire data sets, model accuracies for Confusion Matrices (balanced data set versus entire data set) were significantly lower (one-tailed T-test, $p<0.04$). Similar results were obtained for AUC (one-tailed T-test, $p<0.01$). The presumed reason for these differences is that the minority class was under-represented in the entire data sets. The results demonstrate the need of an alternative approach to improve model predictions, such as modeling using SMOTE approaches.

SMOTED data sets. Validation data sets (not used in training or SMOTED) were employed to assess prediction accuracies of the SMOTED models. Table 4 shows that the prediction accuracies based on the Confusion Matrices were reasonable for the SWL and URS models ranging from 82.6% to 93.0%. Interestingly, accuracies based on AUC were sub-optimal, with prediction values ranging from 49.6% to 70.5%. This finding suggests AUCs are more sensitive to the number of minority records (and/or the noise) in the validation data sets than the Confusion Matrices. We will investigate this issue in the next section below.

Comparison of the prediction accuracies of the models (two-tailed T-tests) using the validation data sets and the entire data sets revealed no significant differences for the Confusion Matrix or AUC results (Table 4, Tables S9-S16, Figures S3-S4). The significance of this finding is that models trained with the SMOTED data sets yielded relatively consistent outcomes regardless of the data sets used to test them. Of note, the SMOTED data sets were not used to test the models – they were only used to train the models.

Predictions based on ensembled ANN models. Ensemble processing was used to improve upon model predictions and assess the variability of the predictions of each patient record. This was accomplished by calculating the averages and standard deviations of the predictions from 10 independently SMOTED ANN models. The averaged values were then used to assess model performance (Confusion matrix and AUC).

Model accuracies were 85.0% for SWL stone removal results based on the Confusion Matrix (Table 5) and 74.8% for results based on AUC (Figure 1A). Model accuracy was 95.1% for SWL treatment complication results based on the Confusion Matrix (Table 6) and 66.3% for those based on AUC (Figure 1B).

Model accuracy was 91.2% for URS stone removal results based on the Confusion Matrix (Table 7) and 77.2% for those based on the AUC (Figure 1C), suggesting moderate to high precision.

Model accuracy was 93.2% for URS treatment complication results based on the Confusion matrix (Table 8) and 78.9% for results based on AUC (Figure 1D).

The model accuracies for the averaged SMOTED ANN models based on the entire data sets are summarized in Table 9. Two-way T-tests showed no significant differences in predicted outcomes based on Confusion Matrices of individually trained ANN models (Table 4) and those of averaged ANN models (Table 9). However, there were significant improvements in predictions based on AUC results ($P<0.027$). Specifically, the averaged AUC values increased from 58.0% to 73.4%, suggesting that noise in the data was responsible for the substantially lower AUC results previously reported (Table 4).

In summary, ensemble models improved predictions in two ways: (i) it significantly improved AUC results, and (ii) it enabled Users to access the precision of predictions; those having low standard deviations versus those with high standard deviations, which is important for making intervention decisions of individual patients with kidney stones based on the KSDE.

Assessment of KSDE Performance.
Overall

The incorrect KSDE predictions could be separated into two categories: (i) those within one standard deviation of the actual value, and (ii) those outside the standard deviation. Incorrect predictions in the first category ranged from 1.1% to 6.1% of the total depending on intervention and outcome, while those in the second ranged from 2.6% to 8.4% (Table 10). Combining the number of correct predictions with the incorrect predictions in the first category revealed that the KSDE was reasonably accurate with values ranging from 91.5% to 97.4% (Table 10).

Individual patients
Since the KSDE was designed to predict outcomes for individual patients, predictions of 10 randomly selected individual patient records were compared to corresponding actual values in the SWL and URS data sets (Table 11) and the suggested intervention was determined.

SWL stone removal and treatment complications
All actual values for SWL stone removal indicate that the stones were <4 mm after treatment. The KSDE correctly predicted 9 records were <0.5. One of the records was >0.5 but also had a large standard deviation, indicating the prediction was within one standard deviation of the correct answer. The predictions represent 10 of the 13848 records (91.5%) shown in Table 10.

Nine of the 10 actual records for SWL treatment complications were ‘0’, indicating no treatment complications, but one record was ‘1’ indicating a treatment complication. The KSDE correctly predicted 9 of 10 records but one treatment (i.e., SWL 4) was predicted as a treatment complication with high standard deviation. The significance of this finding is the prediction has high uncertainty but within one standard deviation of the correct answer. The correct predictions are represented as 9 for the 14379 records (95.1%) shown in Table 10 and the

uncertain one represents 162 of the 15126 records (1.1%) that are classified as incorrect but within one standard deviation of the correct prediction.

Six of 10 suggested interventions were categorized, as "SWL or URS" because SWL and URS predicted values were <0.5. Three of the suggested interventions were URS (only) because the scoring system showed that SWL was greater than URS (data not shown). One of the suggested interventions was SWL (only) because the standard deviation of URS treatment complication prediction was >0.25 (data not shown).

URS stone removal and treatment complications

Nine of the 10 actual records for URS stone removal were '0', indicating successful stone removal, but one record was '1', indicating that the stone was >4mm after treatment. The KSDE correctly predicted 8 of the 10 records. One of the two incorrectly predicted records had a high standard deviation (20%) indicating that the prediction was within one standard deviation of the correct value. This record represents one of the 130 (6.1%) shown in Table 10. The other record was a false negative (in bold) and represents one of the 102 records (4.8%) in Table 10.

All ten actual records for URS treatment complications were '0', indicating no treatment complications and the KSDE correctly predicted these records. These predictions represent 10 of the 1950 records (92.2%) shown in Table 10.

Six of 10 suggested interventions were categorized, as "SWL or URS" because SWL and URS predicted values were <0.5. Three records were categorized as SWL (only) because the scoring system found that URS > SWL (data not shown). One record was categorized as 'URS' because the scoring system found that URS < SWL.

In summary, the KSDE demonstrated reasonable accuracy in predicting outcomes based on patient information. To aid healthcare providers in counseling patients and determining the optimal treatment options for kidney stones, we have developed a user-friendly KSDE web interface, which can be accessed at http://peteranoble.com/webapps.html.

DISCUSSION

The primary motivation of our study was driven by the desire to provide healthcare professionals with a data-driven tool to accurately predict treatment outcomes based on patient information and intervention (SWL and URS). To our knowledge, this is the first large-scale study to predict kidney stone outcomes using ANN modeling. Our study is unique from other studies because the interventions took place at multiple institutions (n=41+) by different medical professionals (n=41+) using a variety of SWL and URS instruments (Table 1 and 2). Hence, the results should be generalizable and not specific to a particular institution, healthcare professional, or instrument.

The secondary motivation was to demonstrate the utility of ANN models to solve complex healthcare problems. Our initial studies using balanced data sets yielded sub-optimal results (Table 3), presumably due to the minority class biasing the predictions when tested with the entire data sets. The SMOTED data substantially increased the representation of the minority class and consequently improved predictions, as shown in this study and others [20-23]. Ensembling by averaging the predictions of multiple diverse models, reduced the error and improved upon the final predictions (compare Table 4 to Table 9). The diverse models in our study were due to different random splits of the data, randomization of the SMOTE process, and randomization of the initial sets of weights and biases of the ANN models prior to training. Previous studies have used ensemble processes to improve predictions over those made by individually trained models [24,25]. An additional advantage of the ensemble process in our study was that the variability of the predictions for individual patient records could be determined.

Model predictions based on Confusion Matrix versus those on AUC. We investigated prediction accuracies using Confusion Matrices and AUCs to highlight similarities and differences of the two assessment approaches.

A Confusion Matrix measures the performance of a classifier using a fixed threshold. Predictions <0.5, for example, were classified as '0', which corresponds to either 'successful stone removal' or 'no treatment complications', and predictions >0.5 were classified as '1', which corresponds to 'stone removal failure' or 'treatment complications'. The accuracy of a model was defined by the sum of the True Positives and True Negatives divided by the total number of samples and reported as a percent.

In contrast, AUC examines the performance of a classifier without any fixed threshold -- every possible threshold is examined and plotted as a point on the curve -- and it is reported as a percent. The two approaches differ because AUC is apparently more sensitive to noise in the data than the Confusion Matrix, as demonstrated in this study by the improvement of AUC values after multiple independent predictions were ensembled.

Input variables to the KSDE. Previous studies have shown SWL variables affecting treatment outcomes include gender [26-29], age [26,29-31], SSD [27,32-37], BMI [26,37], stone number [26,29,30], stone size [26-30,33,34,37-43], stone location [26,28-30,35,38,39,43,44], and stone characteristics [26,27,30-35,38,41-45]. Variables affecting URS outcomes include stone number [46,47], stone size [40,46], stone location [46,47], and stone characteristics [46,47].

While some overlap exists in the variables affecting SWL and URS outcomes, more variables have been shown to affect SWL outcomes than URS outcomes. These differences were considered during the construction of the KSDE and explain the different number of input variables used to predict SWL and URS outcomes in our study. The choice of input variables was also dependent on the number of missing or erroneous values (e.g., BMI of >100) in the data sets since rows and columns containing numerous missing or erroneous values were excluded from the study.

Comparison to other studies in the literature

Nomograms and mathematical models have been used to predict SWL and URS outcomes in many previous studies. Nomograms are graphical decision-making tools that are easy to use, and they do not require knowledge of the underlying equation that the nomogram represents. Predictive mathematical models consist of coefficients that are multiplied by input variables and summed to yield a predictive outcome. ANN models fall into this category with the coefficients being the weights and biases of the trained network.

Here, we briefly document previous studies (in chronological order by intervention) and where appropriate, mention their limitations.

SWL studies

Kanao et al. [48] created one of the first nomograms to predict stone-free rates based on 435 patients. While the nomogram considered stone size, stone location, and stone number, critics argued that their approach was inadequate [49] because it did not consider stone density and skin-to-stone distance (SSD).

Vakalopoulos et al. [26] constructed a mathematical model predicting the successful outcomes of 1712 patients. The approach was unique from others because the equations were presented. The stated limitations are: (i) different stone locations (i.e., renal, ureter, and total) required different models; (ii) the models would have to be adjusted for different lithotripters; and (iii), the model needed to be validated prospectively to prove its usefulness.

Two studies developed nomograms predicting SWL stone-free rates in children. The Onal et al. [50] model was based on 395 patients. The limitation of the study was that the model was based on one urologist at a single institution, and a single instrument and the approach has not been externally validated. The Dongan and Tekgul [51] predicted stone-free rates and complication rates. Yanaral et al. [52] argued that both Onal et al. [50] and Dongan and Tekgul [51] studies could be improved by the addition of variables such as stone density, degree of obstruction, shock power, and number of shocks applied.

Wiesenthal et al. [49] examined 422 patients to find that predictors of successful lithotripsy differed by stone location and therefore developed two mathematical equations: one for the kidney and the other for the ureter. The stated limitations are that the models did not consider the different types of lithotripters, nor did they include a diversity of institutions and operators.

Tran et al. [45] developed the Triple D score to predict stone-free rates in 235 patients. The model was developed by applying threshold values to AUC curves for ellipsoid stone volume, SSD, and stone density. The score was based on the sum of the number of parameters that fell below the thresholds. The research has been validated by Ichiyanagi et al. [53] with 226 patients.

Kim et al. [44] predicted stone-free rates for 3028 patients from three independent institutions and developed a nomogram based on sex, stone location, stone number, stone size, mean Hounsfield unit and grade of hydronephrosis. The model could also be used to advise patients on the likelihood of single or multiple SWL treatments.

Ickiyanagi et al. [53] developed the Quadruple D score based on 226 patients to predict renal stone free status. The scoring system was defined as the sum of the Triple D score [43] and a number based on stone location in the kidney. The stated limitations were: (i) the score did not consider stone morphology or hydronephrosis grade; (ii) the score was not tested on stones in the ureters; (iii) the study had limited diversity as it was based on Japanese patients; and (iv) it has not been externally validated.

Yoshioka et al. [43] developed an integer score-based prediction model (S3HoCKwave score) for assessing SWL failure based on 2271 patients. The study was conducted at several medical centers and was shown to be superior to Triple D score developed by Tran et al. [45]. In the model, continuous outcomes were converted to dichotomous outcomes, and then multivariable logistic regression analysis calculated the coefficients for each prediction. The values of each prediction were rounded, multiplied by 10 and summed. Assessment of performance was based on internal and external validation. The stated limitations are that the study was based on Asian population and limited to non-contrast-enhanced computed tomography.

URS studies

Resorlu et al. [46] developed a scoring system to predict stone free status based on 207 patients using the following variables: stone size, composition, stone number, renal malformation and lower pole infundibulopelvic angle. Each variable (excluding composition) was scored as either zero or one based on yes or no answers. While the system was limited to a few patients, it has been externally verified by Wang et al. [54] and Bozkurt et al. [55].

Imamura et al. [56] developed a nomogram based on 412 patients that predicted stone free rate. De Nunzi et al. [57] validated the Imamura nomogram using 275 European patients.

Jung et al. [58] developed a modified S-ReSC score based on 88 patients to predict stone free status; but the low number of patients limits the usefulness of the score although it has been externally evaluated [55].

Ito et al. [47] develop a scoring system for stone free status based on 310 patients using stone volume, stone location, operator experience, stone number and presence of hydronephrosis. The score was derived by the sum of individual scores. The stated limitation of the system is too few patients but it has been externally evaluated [55].

Xiao et al. [59] developed the R.I.R.S system based on 382 patients to predict stone free status of 4 parameters: renal stone density, inferior pole stone, renal infundibular length and cumulative stone diameter. It has been externally evaluated [55].

Bozkurt et al. [55] examined four of the five URS nomograms mentioned above [i.e., 46,47,58,59] with 949 patients from two institutions. While the nomograms predicted stone free status and treatment complications with varying degrees of success, Bozkurt stated that the nomograms have limitations, and an ideal system has yet to be developed.

CONCLUSIONS

The strengths of our study are that the models were based on 17242 patients – far more than other studies; and the predictions should be generalizable because the data were collected from many different institutions, with different healthcare professionals, and a variety of SWL and URS instruments. One limitation of our study is its retrospective design, which may have led to biases and reduced the predictive accuracies of the ensembled models. Ongoing prospective studies may improve upon our findings.

This is the first large-scale multi-site study to develop a KSDE that accurately predicts SWL and URS outcomes for prospective patients. A practical outcome of this research is a KSDE web interface that can help healthcare providers in counseling patients and determining the optimal treatment options: http://peteranoble.com/webapps.html.

Acknowledgements: We thank Dr. Alex Pozhitkov from the City of Hope Cancer Research Center for his critical comments in an earlier version of the manuscript and testing the KSDE web interface. We also thank Derek Soetemans and Ifeanyi Okwuchi from the Focus21 team for their help in critical improvements in earlier versions of the manuscript.

Funding. This study was supported by funds from Translational Analytics and Statistics.

Authors Contributions. The project was conceived and designed by PAN. PAN analyzed the data, wrote the paper, and built the KSDE web interface. All authors read and approved the contents of the paper.

Ethic Statement. Individual patient consent was not required as no patient identifiable records were used in the study.

Data Availability. The data sets were obtained from the Kidney Stone Registry (http://kidneystoneregistry.com.s3-website-us-west-2.amazonaws.com/).

Supplementary Materials. The file **'**Supplementary Materials.docx' contains 16 tables and 4 figures.

Conflict of Interest. The authors declare no conflict of interest.

Abbreviations:

ANN	Artificial Neural Network

AUC	Area-Under-the-Curve
BMI	Body Mass Index
DCD2	Dornier Compact Delta II
DCD3	Dornier Compact Delta III
DCS	Dornier Compact Sigma
DMH30	Dornier Medilas H30
DMH35	Dornier Medilas H35
SWL	Extracorporeal shock wave lithotripsy
LV100	Lumenis Versapulse 100 watt
LV20	Lumenis Versapulse 20 watt
NIH	National Institute of Health
OC30	Odyssey Convergent 30 watts
KSDE	Kidney Stone Decision Engine
SF2	Storz F2
SMOTE	Synthetic Minority Oversampling Technique
SSLXT	Storz SLX-T
URS	Ureterorenoscopy

REFERENCES

1. Romero V, Akpinar H, Assimos DG. Kidney stones: a global picture of prevalence, incidence, and associated risk factors. Rev Urol. 2010 Spring;12(2-3):e86-96. PMID: 20811557; PMCID: PMC2931286.
2. Scales CD Jr, Smith AC, Hanley JM, Saigal CS; Urologic Diseases in America Project. Prevalence of kidney stones in the United States. Eur Urol. 2012 Jul;62(1):160-5. doi: 10.1016/j.eururo.2012.03.052. Epub 2012 Mar 31. PMID: 22498635; PMCID: PMC3362665.
3. Scales CD Jr, Tasian GE, Schwaderer AL, Goldfarb DS, Star RA, Kirkali Z. Urinary Stone Disease: Advancing Knowledge, Patient Care, and Population Health. Clin J Am Soc Nephrol. 2016 Jul 7;11(7):1305-12. doi: 10.2215/CJN.13251215. Epub 2016 Mar 10. PMID: 26964844; PMCID: PMC4934851.
4. Joshi HB, Johnson H, Pietropaolo A, Raja A, Joyce AD, Somani B, Philip J, Biyani CS, Pickles T. Urinary Stones and Intervention Quality of Life (USIQoL): Development and Validation of a New Core Universal Patient-reported Outcome Measure for Urinary Calculi. Eur Urol Focus. 2021 Jan 8:S2405-4569(20)30313-8. doi: 10.1016/j.euf.2020.12.011. Epub ahead of print. PMID: 33423970.
5. Moudi E, Hosseini SR, Bijani A. Nephrolithiasis in elderly population; effect of demographic characteristics. J Nephropathol. 2017 Mar;6(2):63-68. doi: 10.15171/jnp.2017.11. Epub 2016 Dec 17. PMID: 28491855; PMCID: PMC5418072.
6. Krambeck AE, Lieske JC, Li X, Bergstralh EJ, Melton LJ 3rd, Rule AD. Effect of age on the clinical presentation of incident symptomatic urolithiasis in the general population. J Urol. 2013 Jan;189(1):158-64. doi: 10.1016/j.juro.2012.09.023. Epub 2012 Nov 16. PMID: 23164393; PMCID: PMC3648841.
7. Mains EAA, Blackmur JP, Sharma AD, Gietzmann WK, El-Mokadem I, Stephenson C, Wallace S, Phipps S, Thomas BG, Tolley DA, Cutress ML. Shockwave Lithotripsy Is an Efficacious Treatment Modality for Obese Patients with Upper Ureteral Calculi: Logistic Regression and

Matched-Pair Analyses from a Dedicated Center Comparing Treatment Outcomes by Skin-to-Stone Distance. J Endourol. 2020 Apr;34(4):487-494. doi: 10.1089/end.2019.0717. Epub 2020 Mar 27. PMID: 32030994.
8. Raja A, Hekmati Z, Joshi HB. How Do Urinary Calculi Influence Health-Related Quality of Life and Patient Treatment Preference: A Systematic Review. J Endourol. 2016 Jul;30(7):727-43. doi: 10.1089/end.2016.0110. Epub 2016 May 16. PMID: 27080725.

9. Ramesh S, Chen TT, Maxwell AD, Cunitz BW, Dunmire B, Thiel J, Williams JC, Gardner A, Liu Z, Metzler I, Harper JD, Sorensen MD, Bailey MR. In Vitro Evaluation of Urinary Stone Comminution with a Clinical Burst Wave Lithotripsy System. J Endourol. 2020 Nov;34(11):1167-1173. doi: 10.1089/end.2019.0873. Epub 2020 Mar 20. PMID: 32103689; PMCID: PMC7698855.
10. Legemate JD, Marchant F, Bouzouita A, Li S, McIlhenny C, Miller NL, Saita A, de la Rosette JJ. Outcomes of Ureterorenoscopic Stone Treatment in 301 Patients with a Solitary Kidney. J Endourol. 2017 Oct;31(10):992-1000. doi: 10.1089/end.2017.0180. Epub 2017 Sep 20. PMID: 28826249.
11. McAteer JA, Evan AP. The acute and long-term adverse effects of shock wave lithotripsy. Semin Nephrol. 2008 Mar;28(2):200-13. doi: 10.1016/j.semnephrol.2008.01.003. PMID: 18359401; PMCID: PMC2900184.
12. Rodr'guez D, Sacco DE. Minimally invasive surgical treatment for kidney stone disease. Adv Chronic Kidney Dis. 2015 Jul;22(4):266-72. doi: 10.1053/j.ackd.2015.03.005. PMID: 26088070.
13. Chew BH, Zavaglia B, Paterson RF, Teichman JM, Lange D, Zappavigna C, Matlaga BR, Nunez-Nateras R, Bruhn A, Altamar HO, Humphreys MR, Shah O, Miller NL. A multicenter comparison of the safety and effectiveness of ureteroscopic laser lithotripsy in obese and normal weight patients. J Endourol. 2013 Jun;27(6):710-4. doi: 10.1089/end.2012.0605. PMID: 23521213.
14. Constanti M, Calvert RC, Thomas K, Dickinson A, Carlisle S. Cost analysis of ureteroscopy (URS) vs extracorporeal shockwave lithotripsy (SWL) in the management of ureteric stones <10 mm in adults: a UK perspective. BJU Int. 2020 Mar;125(3):457-466. doi: 10.1111/bju.14938. Epub 2019 Dec 2. PMID: 31663246.
15. Aboumarzouk OM, Kata SG, Keeley FX, McClinton S, Nabi G. Extracorporeal shock wave lithotripsy (SWL) versus ureteroscopic management for ureteric calculi. Cochrane Database Syst Rev. 2012 May 16;(5):CD006029. doi: 10.1002/14651858.CD006029.pub4. PMID: 22592707.
16. Sarkissian C, Noble M, Li J, Monga M. Patient decision making for asymptomatic renal calculi: balancing benefit and risk. Urology. 2013 Feb;81(2):236-40. doi: 10.1016/j.urology.2012.10.032. PMID: 23374767.
17. National Institute for Health and Care Excellence. 2019. Surgical Treatment Intervention Evidence Review. NICE guidline NG118.
18. Noble, P.A. and E. Tribou. Neuroet: an easy-to-use artificial neural network for ecological and biological modelling. 2007. Ecological Modelling 203:87-98. noble_tribou_2007.pdf
19. Chawla, NV, KW Bowyer, LO Hall and WP Kegelmeyer. 2011. SMOTE: Synthetic minority over-sampling technique. J Artificial Intelligence Research 16: 321-357. chawla_2011.pdf

20. Sakr S, Elshawi R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, Blaha MJ, Al-Mallah MH. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. BMC Med Inform Decis Mak. 2017 Dec 19;17(1):174. doi: 10.1186/s12911-017-0566-6. PMID: 29258510; PMCID: PMC5735871.
21. Muaz A, Jayabalan M, Thiruchelvam V. A comparison of data sampling techniques for credit card fraud detection. 2020. International J. Advanced Computer Science and Applications 11:477-485.
22. Waqar M, Dawood H, Dawood H, Majeed N, Banjar A, Alharbey R. An efficient SMOTE-based Deep Learning model for heart attack prediction. 2020. Scientific Programming Article ID 6621622 https//doi:10.1155/2021/6621622
23. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS One. 2017 Jul 24;12(7):e0179805. doi: 10.1371/journal.pone.0179805. PMID: 28738059; PMCID: PMC5524285.
24. Billings WM, Morris CJ, Della Corte D. The whole is greater than its parts: ensembling improves protein contact prediction. Sci Rep. 2021 Apr 13;11(1):8039. doi: 10.1038/s41598-021-87524-0. PMID: 33850214; PMCID: PMC8044223.
25. Verma S, Sharma N, Singh A, Alharbi A, Alosaimi W, Alyami H, Gupta, D, Goyal N. 2022. An intelligent forecasting model for disease prediction using stack ensembling approach. Computers, Materials and Continua 70: 6041-6055.
26. Vakalopoulos I. Development of a mathematical model to predict extracorporeal shockwave lithotripsy outcome. J Endourol. 2009 Jun;23(6):891-7. doi: 10.1089/end.2008.0465. PMID: 19441881.
27. Shinde S, Al Balushi Y, Hossny M, Jose S, Al Busaidy S. Factors Affecting the Outcome of Extracorporeal Shockwave Lithotripsy in Urinary Stone Treatment. Oman Med J. 2018 May;33(3):209-217. doi: 10.5001/omj.2018.39. PMID: 29896328; PMCID: PMC5971054.
28. Bovelander E, Weltings S, Rad M, van Kampen P, Pelger RCM, Roshani H. The Influence of Pain on the Outcome of Extracorporeal Shockwave Lithotripsy. Curr Urol. 2019 Mar 8;12(2):81-87. doi: 10.1159/000489424. PMID: 31114465; PMCID: PMC6504796.
29. Cui HW, Silva MD, Mills AW, North BV, Turney BW. Predicting shockwave lithotripsy outcome for urolithiasis using clinical and stone computed tomography texture analysis variables. Sci Rep. 2019 Oct 11;9(1):14674. doi: 10.1038/s41598-019-51026-x. PMID: 31604986; PMCID: PMC6788981.
30. Abe T, Akakura K, Kawaguchi M, Ueda T, Ichikawa T, Ito H, Nozumi K, Suzuki K. Outcomes of shockwave lithotripsy for upper urinary-tract stones: a large-scale study at a single institution. J Endourol. 2005 Sep;19(7):768-73. doi: 10.1089/end.2005.19.768. PMID: 16190825.
31. Chiang BJ, Liao CH, Lin YH. The efficacy of extracorporeal shockwave lithotripsy for symptomatic ureteral stones: Predictors of treatment failure without the assistance of computed tomography. PLoS One. 2017 Sep 20;12(9):e0184855. doi: 10.1371/journal.pone.0184855. PMID: 28931028; PMCID: PMC5607160.
32. Erkoc M, Bozkurt M, Besiroglu H, Canat L, Atalay HA. Success of extracorporeal shock wave lithotripsy based on CT texture analysis. Int J Clin Pract. 2021 Nov;75(11):e14823. doi: 10.1111/ijcp.14823. Epub 2021 Sep 15. PMID: 34491588.

33. Cho KS, Jung HD, Ham WS, Chung DY, Kang YJ, Jang WS, Kwon JK, Choi YD, Lee JY. Optimal Skin-to-Stone Distance Is a Positive Predictor for Successful Outcomes in Upper Ureter Calculi following Extracorporeal Shock Wave Lithotripsy: A Bayesian Model Averaging Approach. PLoS One. 2015 Dec 14;10(12):e0144912. doi: 10.1371/journal.pone.0144912. PMID: 26659086; PMCID: PMC4699456.
34. Lee HY, Yang YH, Lee YL, Shen JT, Jang MY, Shih PM, Wu WJ, Chou YH, Juan YS. Noncontrast computed tomography factors that predict the renal stone outcome after shock wave lithotripsy. Clin Imaging. 2015 Sep-Oct;39(5):845-50. doi: 10.1016/j.clinimag.2015.04.010. Epub 2015 Apr 25. PMID: 25975631.
35. Waqas M, Saqib IU, Imran Jamil M, Ayaz Khan M, Akhter S. Evaluating the importance of different computed tomography scan-based factors in predicting the outcome of extracorporeal shock wave lithotripsy for renal stones. Investig Clin Urol. 2018 Jan;59(1):25-31. doi: 10.4111/icu.2018.59.1.25. Epub 2017 Dec 28. PMID: 29333511; PMCID: PMC5754579.
36. Ali Beigi MF, Keivani Hafshejani Z, Aghahoseini M, Shirani M Impact of body mass index on success, complications and failure of extracorporeal shock wave lithotripsy. J Renal Inj Prev. 2019;8(3):221-224. DOI: 10.15171/jrip.2019.41.
37. Bajaj M, Smith R, Rice M, Zargar-Shoshtari K. Predictors of success following extracorporeal shock-wave lithotripsy in a contemporary cohort. Urol Ann. 2021 Jul-Sep;13(3):282-287. doi: 10.4103/UA.UA_155_19. Epub 2021 Jul 14. PMID: 34421266; PMCID: PMC8343291.
38. Nakasato T, Morita J, Ogawa Y. Evaluation of Hounsfield Units as a predictive factor for the outcome of extracorporeal shock wave lithotripsy and stone composition. Urolithiasis. 2015 Feb;43(1):69-75. doi: 10.1007/s00240-014-0712-x. Epub 2014 Aug 20. PMID: 25139151.
39. Nielsen TK, Jensen JB. Efficacy of commercialised extracorporeal shock wave lithotripsy service: a review of 589 renal stones. BMC Urol. 2017 Jul 27;17(1):59. doi: 10.1186/s12894-017-0249-8. PMID: 28750620; PMCID: PMC5532761.
40. Fankhauser CD, Hermanns T, Lieger L, Diethelm O, Umbehr M, LuginbŸhl T, Sulser T, MŸntener M, Poyet C. Extracorporeal shock wave lithotripsy versus flexible ureterorenoscopy in the treatment of untreated renal calculi. Clin Kidney J. 2018 Jun;11(3):364-369. doi: 10.1093/ckj/sfx151. Epub 2018 Jan 25. Erratum in: Clin Kidney J. 2018 Apr 18;12(2):309-310. PMID: 29992018; PMCID: PMC6007408.
41. Yoon JH, Park S, Kim SC, Park S, Moon KH, Cheon SH, Kwon T. Outcomes of extracorporeal shock wave lithotripsy for ureteral stones according to SWL intensity. Transl Androl Urol. 2021 Apr;10(4):1588-1595. doi: 10.21037/tau-20-1397. PMID: 33968647; PMCID: PMC8100855.
42. Snicorius M, Bakavicius A, Cekauskas A, Miglinas M, Platkevicius G, Zelvys A. Factors influencing extracorporeal shock wave lithotripsy efficiency for optimal patient selection. Wideochir Inne Tech Maloinwazyjne. 2021 Jun;16(2):409-416. doi: 10.5114/wiitm.2021.103915. Epub 2021 Feb 24. PMID: 34136039; PMCID: PMC8193744.
43. Yoshioka T, Ikenoue T, Hashimoto H, Otsuki H, Oeda T, Ishito N, Watanabe R, Saika T, Araki M, Fukuhara S, Yamamoto Y; Okayama-Ehime S.W.L. Study Group. Development and validation of a prediction model for failed shockwave lithotripsy of upper urinary tract calculi using computed tomography information: the S3HoCKwave score. World J Urol. 2020

Dec;38(12):3267-3273. doi: 10.1007/s00345-020-03125-y. Epub 2020 Feb 22. PMID: 32088747; PMCID: PMC7716893.
44. Kim JK, Ha SB, Jeon CH, Oh JJ, Cho SY, Oh SJ, Kim HH, Jeong CW. Clinical Nomograms to Predict Stone-Free Rates after Shock-Wave Lithotripsy: Development and Internal-Validation. PLoS One. 2016 Feb 18;11(2):e0149333. doi: 10.1371/journal.pone.0149333. PMID: 26890006; PMCID: PMC4758663.
45. Tran TY, McGillen K, Cone EB, Pareek G. Triple D Score is a reportable predictor of shockwave lithotripsy stone-free rates. J Endourol. 2015 Feb;29(2):226-30. doi: 10.1089/end.2014.0212. Epub 2014 Sep 19. PMID: 25046472.
46. Resorlu B, Unsal A, Gulec H, Oztuna D. A new scoring system for predicting stone-free rate after retrograde intrarenal surgery: the "resorlu-unsal stone score". Urology. 2012 Sep;80(3):512-8. doi: 10.1016/j.urology.2012.02.072. Epub 2012 Jul 26. PMID: 22840867.
47. Ito H, Sakamaki K, Kawahara T, Terao H, Yasuda K, Kuroda S, Yao M, Kubota Y, Matsuzaki J. Development and internal validation of a nomogram for predicting stone-free status after flexible ureteroscopy for renal stones. BJU Int. 2015 Mar;115(3):446-51. doi: 10.1111/bju.12775. Epub 2014 Aug 13. PMID: 24731157.
48. Kanao K, Nakashima J, Nakagawa K, Asakura H, Miyajima A, Oya M, Ohigashi T, Murai M. Preoperative nomograms for predicting stone-free rate after extracorporeal shock wave lithotripsy. J Urol. 2006 Oct;176(4 Pt 1):1453-6; discussion 1456-7. doi: 10.1016/j.juro.2006.06.089. PMID: 16952658.
49. Wiesenthal JD, Ghiculete D, Ray AA, Honey RJ, Pace KT. A clinical nomogram to predict the successful shock wave lithotripsy of renal and ureteral calculi. J Urol. 2011 Aug;186(2):556-62. doi: 10.1016/j.juro.2011.03.109. PMID: 21684557.
50. Onal B, Tansu N, Demirkesen O, Yalcin V, Huang L, Nguyen HT, Cilento BG, Erozenci A. Nomogram and scoring system for predicting stone-free status after extracorporeal shock wave lithotripsy in children with urolithiasis. BJU Int. 2013 Feb;111(2):344-52. doi: 10.1111/j.1464-410X.2012.11281.x. Epub 2012 Jun 6. PMID: 22672514.
51. Dogan HS and Tekgul S. 2013 Extracorporeal shock wave lithotripsy: Principles of fragmentation techniques. Pediatric Endourology Techniques. Pp. 257 to 263.
52. Yanaral F, Ozgor F, Savun M, Agbas A, Akbulut F, Sarilar O. Shock-wave Lithotripsy for Pediatric Patients: Which Nomogram Can Better Predict Postoperative Outcomes? Urology. 2018 Jul;117:126-130. doi: 10.1016/j.urology.2018.03.045. Epub 2018 Apr 6. PMID: 29630952.
53. Ichiyanagi O, Fukuhara H, Kurokawa M, Izumi T, Suzuki H, Naito S, Nishida H, Kato T, Tsuchiya N. Reinforcement of the Triple D score with simple addition of the intrarenal location for the prediction of the stone-free rate after shockwave lithotripsy for renal stones 10-20 mm in diameter. Int Urol Nephrol. 2019 Feb;51(2):239-245. doi: 10.1007/s11255-018-02066-1. Epub 2019 Jan 2. PMID: 30604235.
54. Wang C, Wang S, Wang X, Lu J. External validation of the R.I.R.S. scoring system to predict stone-free rate after retrograde intrarenal surgery. BMC Urol. 2021 Mar 4;21(1):33. doi: 10.1186/s12894-021-00801-y. PMID: 33663459; PMCID: PMC7934254.
55. Bozkurt IH, Karakoyunlu AN, Koras O, Celik S, Sefik E, Cakici MC, Degirmenci T, Imamoglu MA. External validation and comparison of current scoring systems in retrograde intrarenal

surgery: Multi-institutional study with 949 patients. Int J Clin Pract. 2021 Jun;75(6):e14097. doi: 10.1111/ijcp.14097. Epub 2021 Feb 28. PMID: 33619879.
56. Imamura Y, Kawamura K, Sazuka T, Sakamoto S, Imamoto T, Nihei N, Suzuki H, Okano T, Nozumi K, Ichikawa T. Development of a nomogram for predicting the stone-free rate after transurethral ureterolithotripsy using semi-rigid ureteroscope. Int J Urol. 2013 Jun;20(6):616-21. doi: 10.1111/j.1442-2042.2012.03229.x. Epub 2012 Nov 19. PMID: 23163835.
57. De Nunzio C, Bellangino M, Voglino OA, Baldassarri V, Lombardo R, Pignatelli M, Tema G, Berardi E, Cremona A, Tubaro A. External validation of Imamura nomogram as a tool to predict preoperatively laser semi-rigid ureterolithotripsy outcomes. Minerva Urol Nefrol. 2019 Oct;71(5):531-536. doi: 10.23736/S0393-2249.18.03243-5. Epub 2018 Dec 14. PMID: 30547902.
58. Jung JW, Lee BK, Park YH, Lee S, Jeong SJ, Lee SE, Jeong CW. Modified Seoul National University Renal Stone Complexity score for retrograde intrarenal surgery. Urolithiasis. 2014 Aug;42(4):335-40. doi: 10.1007/s00240-014-0650-7. Epub 2014 Mar 13. PMID: 24623504.
59. Xiao Y, Li D, Chen L, Xu Y, Zhang D, Shao Y, Lu J. The R.I.R.S. scoring system: An innovative scoring system for predicting stone-free rate following retrograde intrarenal surgery. BMC Urol. 2017 Nov 21;17(1):105. doi: 10.1186/s12894-017-0297-0. PMID: 29162070; PMCID: PMC5696735.

TABLES

Table 1. Descriptive statistics for the SWL data set. %, proportion in category; *n*, number in category.

Category	Item	SWL data set (*n*=15126)
Instrument used	Dornier Compact Delta II	6.6% (*n*=1003)
	Dornier Compact Delta III	3.5% (*n*=536)
	Dornier Compact Sigma	3.1% (*n*=472)
	Storz F2	30.7% (*n*=4651)
	Storz SLX-T	56.0% (*n*=8464)
Stone Location	Ureters	21.6% (n=3271)
	Kidney	77.8% (*n*=11771)
	0ther locations	<1.0% (*n*=84)
	Stone side (Left=0, Right=1)	55.4% (*n*=8380)
Stone properties	Stone Width (mm)	8.2 ± 4.4
	Stone Length (mm)	8.7 ± 4.7
Patient Information	Anticoagulants (True=0; False=1)	93.8% (*n*=14192)
	Gender (Male=1, Female=0)	55.1% (n=8338)
	BMI (kg/m^2)	30.1 ± 6.9
	Age at time of procedure (years)	57.0 ± 14.9
	Medical Condition (True=1, False=0)	7.8% (*n*=1173)
Treatment Outcomes	Treatment Complications (False=0; True=1)	4.8% (*n*=732)
	Stone Removal (Success=0; Failure=1)	15.6% (*n*=2353)

Table 2. Descriptive statistics for URS data. %, proportion in category; *n*, number in category.

Category	Item	URS data set (*n*=2116)
Instrument used	Dornier Medilas H20	26.8% (*n*=568)
	Dornier Medilas H30	3.5% (*n*=75)
	Dornier Medilas H35	2.6% (*n*=56)
	Lumenis Versapulse 100 watt	29.3% (*n*=621)
	Lumenis Versapulse 20 watt	34.0% (*n*=723)
	Odyssey Convergent 30 watt	3.4% (*n*=73)
Patient Information		
	Gender (Male=1; Female=0)	54% (*n*=1142)
	Age (years)	56.5 ± 15.5
	BMI (kg/m^2)	30.4 ± 7.7
Treatment Outcomes	Treatment Complications (False=0; True=1)	5.3% (*n*=113)
	Stone Removal (Success=0; Failure=1)	7.2% (*n*=152)

Table 3. Summary of ANN models developed with balanced datasets and tested on the entire data set. Model accuracies were assessed using a Confusion Matrix and AUC. The Confusion Matrices and AUCs are shown in Tables S1-S8 and Figures S1-S2.

Treatment	Predicted outcome	Total records in balanced data set (50%, y=0; 50%, y=1)	Model accuracy (%) with balanced data set (70% training: 30% testing/validation)		Model accuracy (%) with entire SWL data set (*n*=15126)		Model accuracy (%) with entire URS data set (*n*=2116)	
			Confusion matrix	AUC	Confusion matrix	AUC	Confusion matrix	AUC
SWL	Stone removal	4706	72.4	78.5	68.1	74.2	-	-
	Treatment complications	1464	74.9	77.3	51.0	64.1	-	-
URS	Stone removal	304	92.8	95.9	-		16.0	55.1
	Treatment complications	226	80.1	78.6	-		63.1	51.9

Table 4. Summary ANN models developed with SMOTED datasets and tested on the validation data set (hold out) and the entire data set. Model accuracies were assessed using a Confusion Matrix and AUC. The Confusion matrices and AUCs are shown in Tables S9-S16 and Figures S3-S4.

Treatment	Predicted outcome	Model accuracy (%) SWL validation data sets (*n*=4539)		Model accuracy (%) using URS validation data sets (*n*=636)		Model accuracy (%) with entire SWL data set (*n*=15126)		Model accuracy (%) with entire URS data set (*n*=2116)	
		Confusion matrix	AUC	Confusion matrix	AUC	Confusion matrix	AUC	Confusion matrix	AUC
SWL	Stone removal	82.6	66.5	-	-	84.0	70.5	-	-
	Treatment complications	92.2	50.9	-	-	93.0	59.3	-	-
URS	Stone removal	-	-	89.4	64.8	-	-	92.8	55.1
	Treatment complications	-	-	88.2	49.6	-	-	92.6	56.4

Table 5. Confusion Matrix based on averaged predictions of ten ANN models trained on the SMOTED SWL stone removal data set and tested with the entire data set. 0, stone removal success; 1, stone removal failure.

Actual (below) /Predictions (across)	0	1	Sum
0	82.5% (*n*=12475)	2.0% (*n*=298)	12773
1	13.0% (*n*=1967)	2.6% (*n*=386)	2353
			85.0% (*n*=15126)

Table 6. Confusion Matrix based on averaged predictions of ten ANN models trained on the SMOTED SWL treatment complication data set and tested with the entire data set. 0, no treatment complication; 1, treatment complication.

Actual (below) /Predictions (across)	0	1	Sum
0	94.9% (*n*=14355)	0.3% (*n*=39)	14394
1	4.7% (*n*=708)	0.2% (*n*=24)	732
			95.1% (*n*=15126)

Table 7. Confusion Matrix based on averaged predictions of ten ANN models trained on the SMOTED URS stone removal data set and tested with the entire data set. 0, successful stone removal; 1, stone removal failure.

Actual (below) /Predictions (across)	0	1	Sum
0	89.9% (*n*=1902)	2.9% (*n*=62)	1964
1	5.9% (*n*=125)	1.3% (*n*=27)	152
			91.2% (*n*=2116)

Table 8. Confusion Matrix based on averaged predictions of ten ANN models trained on the SMOTED URS treatment complication data set and tested with the entire data set. 0, no treatment complication; 1, treatment complication.

Actual (below) /Predictions (across)	0	1	Sum
0	92.2% (*n*=1950)	2.5% (*n*=53)	2003
1	4.3% (*n*=90)	1.1% (*n*=23)	113
			93.2% (*n*=2116)

Table 9. Summary of model accuracies for ensembled SMOTED ANN models (*n*=10) tested with entire data sets. Model accuracies were assessed using Confusion Matrix and AUC.

Treatment	Predicted outcome	Model accuracy (%) with entire SWL data set (*n*=15126)		Model accuracy (%) with entire URS data set (*n*=2116)	
		Confusion matrix	AUC	Confusion matrix	AUC
SWL	Stone removal	85.0	74.8	-	-
	Treatment complications	95.0	66.3	-	-
URS	Stone removal	-	-	91.2	77.2
	Treatment complications	-	-	93.2	78.9

Table 10. Prediction performance of KSDE (40 model equations) by intervention and outcome.

Intervention (across)	SWL (*n*=15126 records)		URS (*n*=2116 records)	
Outcome	Stone removal	Treatment complications	Stone removal	Treatment complications
Correct predictions	85.0% (*n*=12861)	95.1% (*n*=14379)	89.0% (*n*=1884)	92.2% (*n*=1950)
Incorrect predictions but within STD	6.5% (*n*=987)	1.1% (*n*=162)	6.1% (*n*=130)	5.1% (*n*=110)
Incorrect Prediction	8.4% (*n*=1278)	3.9% (*n*=585)	4.8% (*n*=102)	2.6% (*n*=56)
Correct predictions and/or incorrect predictions within STD	91.5%	96.1%	95.1%	97.4%

Table 11. Ten random selected examples of the prediction performance of KSDE by intervention, outcome and suggested intervention. *, Incorrect prediction but within standard deviation (Stdev); Bold, incorrect prediction.

Intervention by individual patient	Actual stone removal (SR) (0=Success; 1=Failure)	Predicted SR ± Stdev	Actual treat complications (TC) (False=0; True=1)	Predicted TC ± Stdev	Suggested Intervention
SWL 1	0	0.05 ± 0.03	0	0.12 ± 0.06	SWL_or_URS
SWL 2	0	0.32 ± 0.28	1	0.68 ± 0.22	URS
SWL 3	0	0.06 ± 0.18	0	0.05 ± 0.16	SWL_or_URS
SWL 4	0	0.48 ± 0.37	0	0.70 ± 0.42*	URS
SWL 5	0	0.26 ± 0.24	0	0.18 ± 0.23	SWL_or_URS
SWL 6	0	0.06 ± 0.17	0	0.06 ± 0.17	SWL_or_URS
SWL 7	0	0.01 ± 0.14	0	0.28 ± 0.35	SWL_or_URS
SWL 8	0	0.03 ± 0.13	0	0.01 ± 0.20	SWL
SWL 9	0	0.07 ± 0.13	0	0.06 ± 0.16	SWL_or_URS
SWL 10	0	0.68 ± 0.39*	0	0.07 ± 0.21	URS
URS 1	0	0.00 ± 0.14	0	0.18 ± 0.29	URS
URS 2	0	**0.76 ± 0.24**	0	0.02 ± 0.18	SWL
URS 3	0	0.10 ± 0.18	0	0.01 ± 0.15	SWL_or_URS
URS 4	0	0.00 ± 0.14	0	0.01 ± 0.13	SWL_or_URS
URS 5	0	0.00 ± 0.13	0	0.00 ± 0.12	SWL_or_URS
URS 6	0	0.47 ± 0.20	0	0.04 ± 0.14	SWL_or_URS
URS 7	0	0.00 ± 0.15	0	0.00 ± 0.12	SWL_or_URS
URS 8	1	0.37 ± 0.20*	0	0.27 ± 0.29	SWL
URS 9	0	0.44 ± 0.20	0	0.27 ± 0.29	SWL
URS 10	0	0.19 ± 0.20	0	0.01 ± 0.18	SWL_or_URS

FIGURES

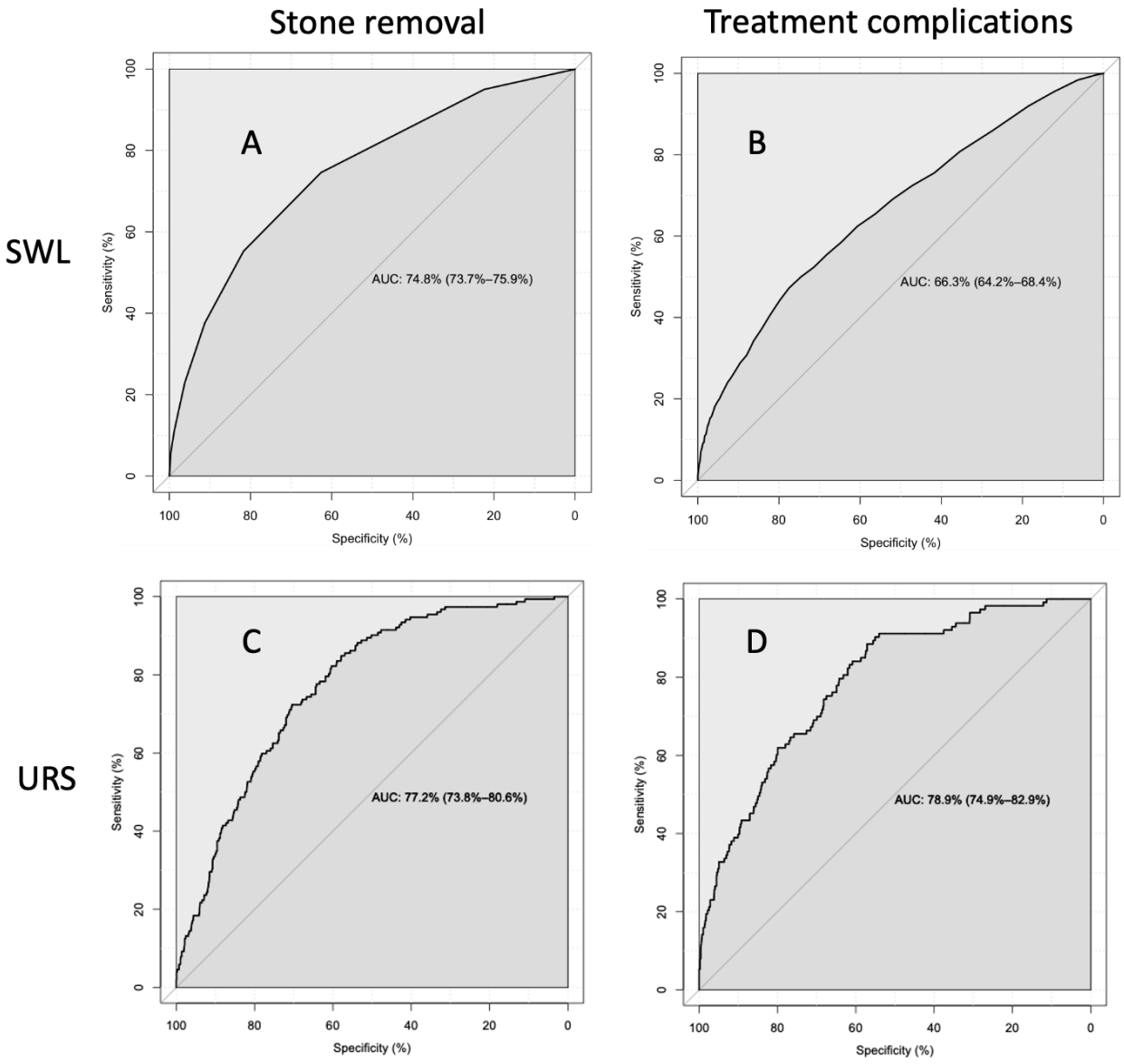

Figure 1. AUCs for averaged predictions from ten ANN models trained on SMOTED SWL stone removal data set (A) and treatment complication (B) and tested with the entire data set (*n*=15126 records). AUC for averaged predictions from ten ANN models trained on SMOTED URS stone removal data set (C) and treatment complication data set (D) and tested with the entire data set (*n*=2116 records).

+