EUROPEAN UROLOGY FOCUS xxx (xxxx) xxx–xxx

available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Stone Disease

Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit

Robert M. Geraghty[a,b,*], *Anshul Thakur*[c], *Sarah Howles*[d], *William Finch*[e], *Sarah Fowler*[f], *Alistair Rogers*[a], *Seshadri Sriprasad*[g], *Daron Smith*[h], *Andrew Dickinson*[i], *Zara Gall*[j], *Bhaskar K. Somani*[k]

[a] *Department of Urology, Freeman Hospital, Newcastle upon Tyne, UK;* [b] *Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, UK;* [c] *Institute of Biomedical Engineering, University of Oxford, Oxford, UK;* [d] *Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK;* [e] *Department of Urology, Norfolk and Norwich University Hospital, Norwich, UK;* [f] *Comparative Audit Service, Royal College of Surgeons of England, London, UK;* [g] *Department of Urology, Dartford and Gravesham NHS Trust, Dartford, UK;* [h] *Institute of Urology, University College Hospital London, London, UK;* [i] *Department of Urology, University Hospitals Plymouth NHS Trust, Plymouth, UK;* [j] *Department of Urology, Stockport NHS Foundation Trust, Stockport, UK;* [k] *Department of Urology, University Hospital Southampton NHS Foundation Trust, Southampton, UK*

Article info

Article history:
Accepted January 21, 2024

Associate Editor: Christian Gratzke

Keywords:
Percutaneous nephrolithotomy
Machine learning
Outcomes
Prediction
Endourology

Abstract

Background and objective: Machine learning (ML) is a subset of artificial intelligence that uses data to build algorithms to predict specific outcomes. Few ML studies have examined percutaneous nephrolithotomy (PCNL) outcomes. Our objective was to build, streamline, temporally validate, and use ML models for prediction of PCNL outcomes (intensive care admission, postoperative infection, transfusion, adjuvant treatment, postoperative complications, visceral injury, and stone-free status at follow-up) using a comprehensive national database (British Association of Urological Surgeons PCNL).

Methods: This was an ML study using data from a prospective national database. Extreme gradient boosting (XGB), deep neural network (DNN), and logistic regression (LR) models were built for each outcome of interest using complete cases only, imputed, and oversampled and imputed/oversampled data sets. All validation was performed with complete cases only. Temporal validation was performed with 2019 data only. A second round used a composite of the most important 11 variables in each model to build the final model for inclusion in the *shiny* application. We report statistics for prognostic accuracy.

Key findings and limitations: The database contains 12 810 patients. The final variables included were age, Charlson comorbidity index, preoperative haemoglobin, Guy's stone score, stone location, size of outer sheath, preoperative midstream urine result, primary puncture site, preoperative dimercapto-succinic acid scan, stone size, and image guidance (https://endourology.shinyapps.io/PCNL_Demographics/). The areas under the receiver operating characteristic curve was >0.6 in all cases.

Conclusions and clinical implications: This is the largest ML study on PCNL outcomes to date. The models are temporally valid and therefore can be implemented in clinical

* Corresponding author. Department of Urology, Freeman Hospital, Freeman Road, Newcastle upon Tyne NE7 7DN, UK.
E-mail address: rob.geraghty@newcastle.ac.uk (R.M. Geraghty).

https://doi.org/10.1016/j.euf.2024.01.011

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

practice for patient-specific risk profiling. Further work will be conducted to externally validate the models.

Patient summary: We applied artificial intelligence to data for patients who underwent a keyhole surgery to remove kidney stones and developed a model to predict outcomes for this procedure. Doctors could use this tool to advise patients about their risk of complications and the outcomes they can expect after this surgery.

1. Introduction

Kidney stone disease is a prevalent and costly condition [1]. Large kidney stones are often treated with percutaneous nephrolithotomy (PCNL) [2]. In addition to the planned outcome of stone removal (ie, stone-free status), PCNL has a number of potential complications, including a need for blood transfusion, postoperative infection, and visceral injury [3]. Several scoring systems have been built in attempts to predict outcomes for individual patients [4]. More recently, (supervised) machine learning (ML) techniques have been used to build models for predicting outcomes of PCNL [5–7]. In comparison to statistical methods, ML can handle highly nonlinear relationships by allowing a computer to predict outcomes on the basis of algorithmic rather than statistical methods (eg, logistic regression). This provides superior accuracy in comparison to traditional statistical methods, especially for rare outcomes. However, the models generated to produce these results can be highly complex and in effect a "black box". To try and demonstrate which variables contribute to a particular outcome, metrics such as Shapley weighting are used to "explain" individual ML predictions [8].

To date, the largest PCNL data set used for ML involved 134 cases [6] and only four outcomes have been described: stone-free status, need for adjuvant treatment, need for stent insertion, and need for blood transfusion. Aminsharifi et al [6] demonstrated that ML using support vector machines had superior accuracy to traditional nomograms (Guy's stone score and the Clinical Research Office of the Endourological Society [CROES] PCNL nomogram).

None of the currently published models have been validated. There are three ways to perform validation, from which prognostic accuracy statistics are generated: internal, external, and temporal validation. Internal validation (often termed the "test" set) simply represents a small subset (usually 20–30%) of the total dataset. External validation uses an external data set. Temporal validation is a form of external validation for which new data are collected from the same source as the training set but in a different (ideally later) time period [9].

To facilitate better personalised prediction of postoperative outcomes, we used a large national database to develop ML models for prediction of seven important PCNL outcomes: stone-free status, need for transfusion, need for intensive care, visceral injury, need for adjuvant treatment, postoperative infection, and postoperative complications. After temporal validation, the best-performing models were used in a web-based application to facilitate individualised predictions.

2. Patients and methods

2.1. Methodology reporting

This study is reported in accordance with the TRIPOD checklist [10], which is included in the Supplementary material.

2.2. Patients and data set

We used data from the British Association of Urological Surgeons (BAUS) PCNL audit, for which the data collection methods have been published [11], but we report them in brief here. Through advertisements at national urological meetings, all surgeons undertaking PCNL in the UK were invited to submit data to the registry using an online interface. An individual record that contained both a unique patient identifier and National Health Service (NHS) number was created for each PCNL procedure. Data were collected between 2014 and 2019 and are detailed below.

2.3. Predictors and outcomes

We built initial models using all 43 preoperative predictive variables within the data set: age; body mass index; preoperative haemoglobin (in g/l); Charlson comorbidity index (score 0–10); age-related Charlson comorbidity index (score 0–11) [12]; number of tracts planned; number of tracts performed; sex; side of stones; previous treatment for urinary tract infection; preoperative antibiotic course; preoperative urine culture; preoperative urine culture result; primary preoperative imaging; secondary preoperative imaging; preoperative dimercapto-succinic acid (DMSA) renogram; catheterisation status; preoperative estimated glomerular filtration rate; prophylactic antibiotics on induction; grade of main operating surgeon; type of anaesthesia; interventional radiologist availability; secondary re-look nephroscopy; stone dimensions (in cm); number of stones; index stone location; other stone location(s); Guy's stone score [13]; maximum Hounsfield units for the index stone on computed tomography (CT) of the kidney, ureter, and bladder; pre-existing nephrostomy tube status; specialty and grade of practitioner performing puncture; puncture site; image guidance for renal puncture; patient position; anatomic placement of tract; size of Amplatz sheath (Fr); type of dilators used; predicted difficulty; accessory procedures; postoperative nephrostomy; and primary and secondary stone extraction techniques.

We examined nine outcomes: visceral injury; need for transfusion; postoperative infection; postoperative complications; need for higher care (high dependency unit [HDU] or intensive therapy unit [ITU]); immediate clearance on intraoperative fluoroscopy; clearance on immediate postoperative imaging (clinician-chosen); stone-free status at follow-up (first outpatient review using radiography, ultrasonography, or CT according to local practice); and need for adjuvant treatment. The following outcomes were also available but were not examined owing to their lack of clinical utility: Clavien-Dindo classification of complications [14], postoperative stay, and survival. However, summary statistics for these outcomes are reported for contexualisation of the database.

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

Following initial model building and testing, we aggregated variable importance (so the online application would be as user friendly as possible, ie, maximal outputs for minimal inputs) for each model and selected the top 12 most important variables to build further models (testing the top 10, 11, and 12 most important variables).

2.4. Sample size calculation

Sample size was calculated for the least likely event, that is, the event for which the largest number of patients was needed (vascular injury necessitating nephrectomy, ~0.1%). Using a 0.1% population proportion, c-statistic of 0.8 (idealised), maximum number of parameters (n=30) and shrinkage of 0.5, a sample size of n = 5609 was calculated [15,16] [see statistical code section 2.11].

2.5. Missing data

For each outcome, cases with missing data for that outcome were excluded. We then built four models for each outcome: complete cases only (further removal of cases with missing data in any column); multiple imputation (using the *mice* [17] package; no collinearity detected before this step); oversampled (using the *ROSE* [18] package); and imputed (multiple) and oversampled. The test sets included only complete cases and were neither imputed nor oversampled.

2.6. Model selection

We used logistic regression (traditional statistical modelling technique), extreme gradient boosting (XGB) [19], and deep neural networks (DNNs) for model building. Explanations of these models are detailed below.

2.6.1. Gradient boosting machine

Gradient boosting machine [20] and one of its variants, gradient tree boosting (GTB), is an ensemble procedure that iteratively fits simple statistical models to the data. GTB uses classification trees as simple statistical models to model the data. Iteratively, GTB evaluates how well the current model performs, adds another tree to the errors made previously, and then updates the model by adding the regression tree to the ensemble. We use XGB [21], one of the most popular implementations of GTB, which allows for fast computation.

2.6.2. Artificial neural network

Inspired by neurons of the human brain [22], an artificial neural network (ANN) is a nonlinear aggregate extension of simpler regression methods. The network transforms all the input information from the predictors, in both a linear and nonlinear fashion, and passes the result to the next layer. This is repeated until an output layer is reached that forms the prediction of the network. ANNs with more than one hidden layer are termed DNNs [23].

2.7. Model building

All models were built in R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria) [24] using base R, *caret* [25], and *keras* [26]. Initial models were built using XGB only to screen for poorly predicted outcomes, with final model building using XGB, logistic regression, and DNN.

2.8. Validation

Data sets were randomly split into a training set (70% of the total) and a test set (30%) using *sample* in R. The test set (complete cases only) was used to internally validate the models for each outcome in the initial model-building round (Fig. 1). For the final round, the data set was segregated by year of data collection (training set, 2014–2018; temporal validation, 2019).

A hyperparameter tuning strategy was designed to optimise the area under the receiver operating characteristic (ROC) curve (AUC). Hyperparameters for the final models are detailed in Supplementary Table 1.

2.9. Statistical analysis

Summary statistics are provided for the overall data set. We report prognostic accuracy statistics following internal and temporal validation for each model: overall accuracy with 95% confidence interval, sensitivity,

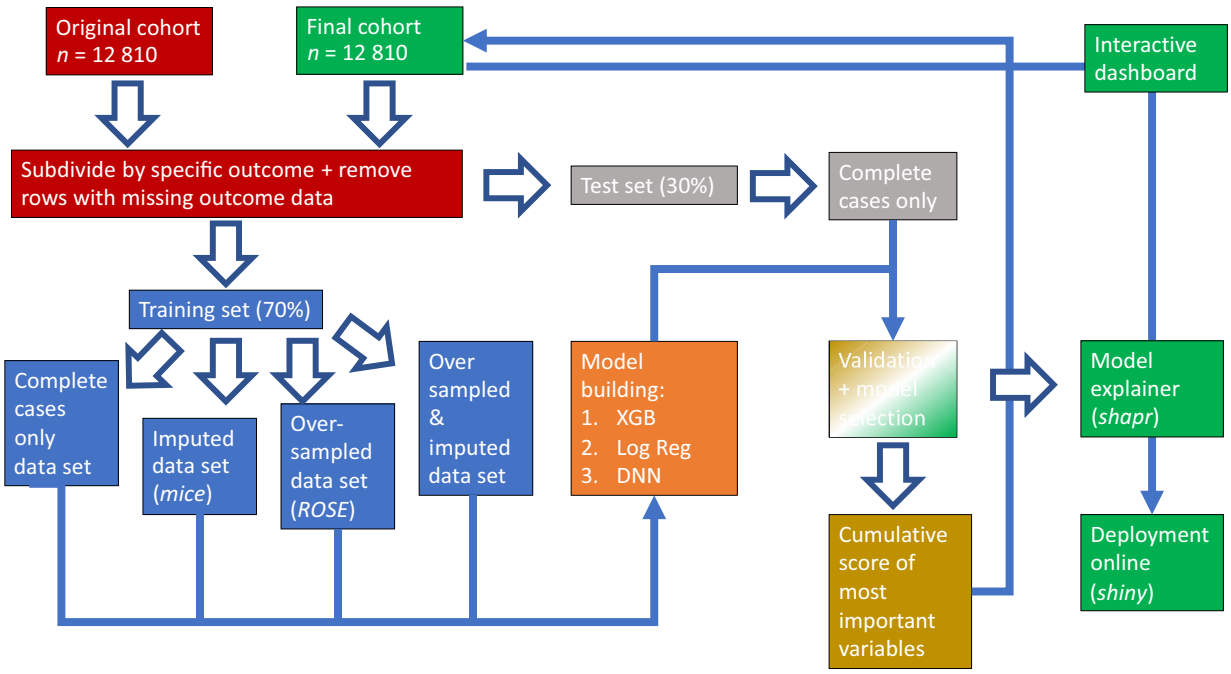

Fig. 1 – Flow diagram of the process steps from collection of raw data to final model use. R packages used at particular points are shown in italic font. XGB = extreme gradient boosting; Log Reg = logistic regression; DNN = deep neural network.

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

specificity, and AUC. Results for the negative predictive value and positive predictive value are available in the Supplementary material. ROC curves for all models and calibration curves were generated using the *MLeval* [27], *caret* [25], *pROC*, and *ggplot2* [28] packages.

2.10. Model selection and use

Models that had high accuracy and AUC (ideally ≥0.7) were used online via the *shiny* package [29], along with a dashboard for data exploration. However, if a particular model achieved high accuracy/AUC but was unable to differentiate between outcomes, then sensitivity and specificity were prioritised. Model explainers were built to display Shapley weights using the *shapr* package. The final model is available online at https://endourology.shinyapps.io/PCNL_Demographics/.

2.11. Code availability

The code is available at https://github.com/rg2u17/PCNL_all_outcomes.

3. Results

3.1. Demographics

The overall data set (n = 12 810), included 5914 women (46%). The mean age was 55.4 yr (±18.6), median body mass index was 28.45 kg/m^2 (interquartile range [IQR] 25.0–33.0), and the median Charlson comorbidity index was 0 (IQR 0–1, range 0–15). A total of 3737 patients (29.2%) had a previous urinary tract infection, 4008 (41.4%) had received preprocedural antibiotics (not including antibiotics at induction, which all patients received), and 10 436 (81.5%) had a preprocedural urine culture.

In the overall data set, outcomes were as follows (percentages calculated based on available outcome data, ie, missing data excluded): 5999 patients (70%) had stone clearance on immediate postoperative imaging, 21 (0.18%) had a visceral injury, 17 (0.16%) died, 246 (2%) had a postoperative transfusion, 1329 (12%) had a postoperative infection, 590 (5%) required ITU/HDU admission, 2191 (74%) were stone-free at follow-up, and 468 (16%) needed adjuvant treatment. In terms of postoperative stay, 52 patients (1%) were discharged on the same day, 886 (20%) had a stay of 1 d, 998 (26%) had a stay of 2 d, and 2482 (56%) had a stay of ≥3 d. The distribution of Clavien-Dindo complications was as follows: 311 patients (7%) had grade I, 343 (8%) had grade II, 93 (2%) had grade IIIa, 63 (1%) had grade IIIb, 13 (0.3%) had grade IVa, and seven (0.2%) had grade V complications. Details of outcomes in the training and test data sets are provided in the Supplementary material.

A data dashboard is included in the online application.

3.2. Model selection and use

Supplementary Table 2 details the top 19 variables (by mean variable importance) following initial model building. Prognostic accuracy data for the final models following internal and temporal validation are provided in the Supplementary material.

The models performing best for prediction of ITU/HDU admission, postoperative infection, postoperative transfusion, visceral injury, postoperative complications, adjuvant treatment, and stone-free status at follow-up (clinician-defined) were used online (Figs. 1 and 2 and Table 1). Stone-free status at follow-up was chosen because of the clinical importance of this metric over immediate clearance on fluoroscopy or clearance on inpatient postoperative imaging. Calibration plots are presented in Figure 3.

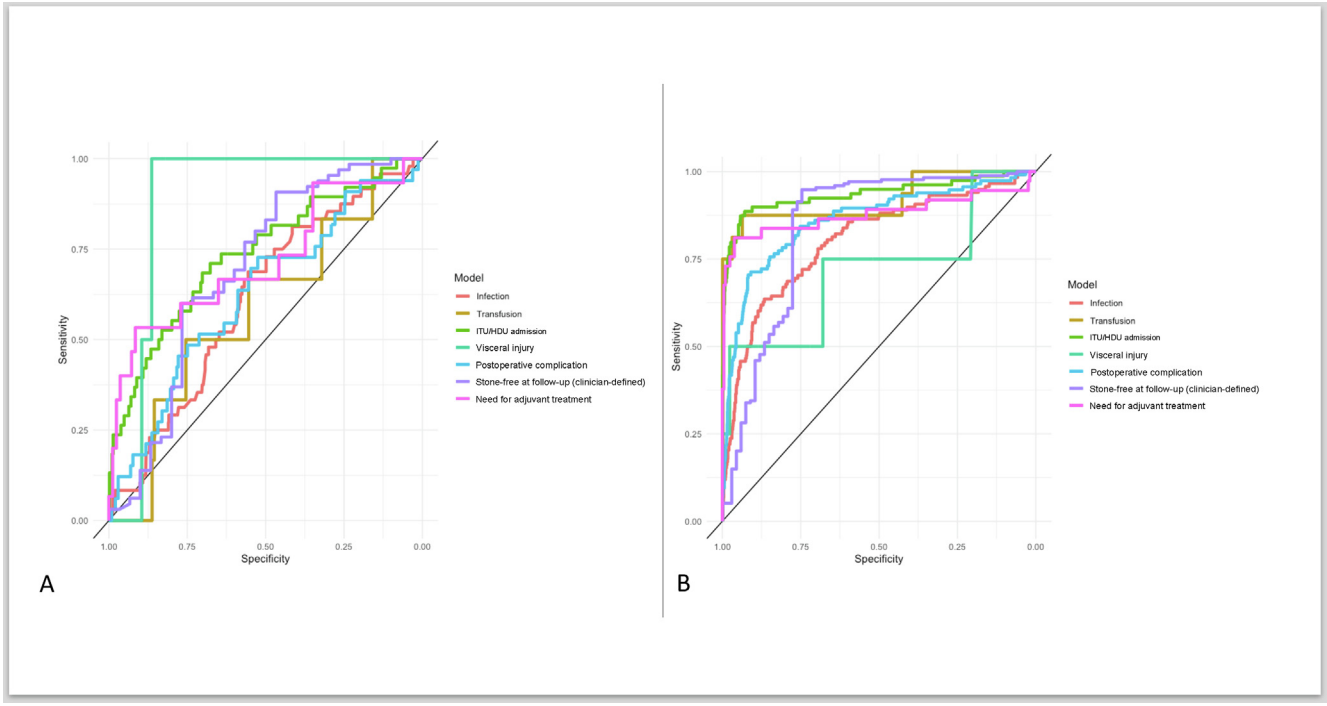

Fig. 2 – Receiver operating characteristic curves for the final models (A) after temporal validation and (B) after internal validation. HDU = high-dependency unit; ITU = intensive therapy unit.

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

Table 1 – Prognostic accuracy statistics based on the test set for the final model for each outcome

Outcome	Final model selected	Training set (outcomes/ total)	Validation set	Test set (outcomes/ total)	Accuracy, % (95% CI)	SSY	SPY	PPV	NPV	AUC (95% CI)
Postoperative infection	XGB_{OI}	6625/13 227	Internal	119/1295	73.0 (70.5–75.4)	0.73	0.73	0.97	0.20	0.82 (0.78–0.86)
			Temporal	48/696	70.1 (66.6–73.5)	0.73	0.38	0.94	0.09	0.59 (0.51–0.67)
Blood transfusion	XGB_{OI}	7416/14 896	Internal	20/1314	98.4 (97.6–99.0)	0.99	0.75	1.00	0.55	0.88 (0.77–0.99)
			Temporal	6/710	91.4 (89.1–93.4)	0.92	0.00	0.99	0.00	0.70 (0.50–0.90)
ITU/HDU admission	XGB_{OI}	6623/13 268	Internal	71/1231	96.0 (94.8–97.0)	0.97	0.78	0.99	0.66	0.94 (0.90–0.97)
			Temporal	38/635	84.3 (81.2–87.0)	0.87	0.47	0.96	0.18	0.74 (0.66–0.82)
Visceral injury	XGB_{OI}	7773/15 508	Internal	3/1352	99.9 (99.6–100.0)	1.00	0.67	1.00	1.00	0.82 (0.47–1)
			Temporal	2/727	99.7 (99.0–99.9)	1.00	0.00	0.99	0.00	0.83 (0.76–0.89)
Postoperative complication	XGB_{OI}	6573/13 124	Internal	156/1308	86.0 (84.0–87.8)	0.88	0.71	0.86	0.44	0.85 (0.81–0.89)
			Temporal	33/711	80.9 (77.8–83.7)	0.83	0.30	0.96	0.08	0.64 (0.54–0.74)
Stone-free at follow-up	XGB_{OI}	1250/2509	Internal	171/242	85.1 (80.0–89.4)	0.77	0.88	0.73	0.90	0.91 (0.87–0.95)
			Temporal	65/95	65.3 (54.8–74.7)	0.40	0.77	0.44	0.74	0.65 (0.53–0.76)
Need for adjuvant treatment	XGB_{OI}	1473/2927	Internal	32/214	93.9 (90.1–96.6)	0.97	0.74	0.96	0.81	0.93 (0.88–0.98)
			Temporal	15/98	73.5 (63.6–81.9)	0.80	0.40	0.88	0.26	0.67 (0.51–0.82)

AUC = area under the receiver operating characteristic curve; CI = confidence interval; HDU = high-dependency unit; ITU = intensive therapy unit; NPV = negative predictive value; PPV = positive predictive value; SPY = specificity; SSY = sensitivity; XGB_{OI} = extreme gradient boosting model, oversampled and imputed.

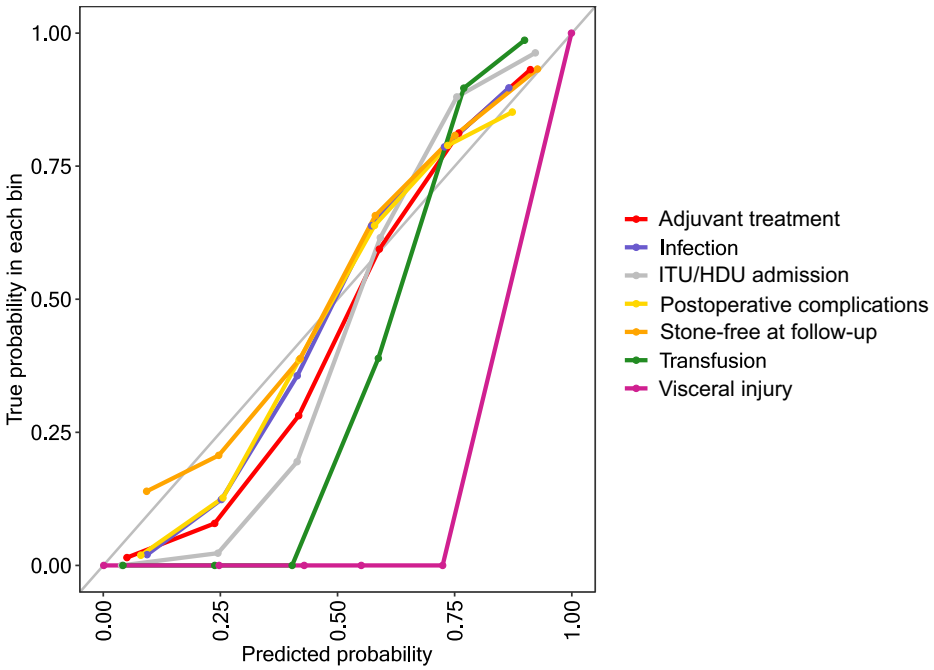

Fig. 3 – Calibration curves for the final models. HDU = high-dependency unit; ITU = intensive therapy unit.

3.3. *Application*

The application is available online at https://endourology.shinyapps.io/PCNL_Demographics/ and is split into four tabs: a Disclaimer (which includes the key), Demographics, ML Predictions, and ML Explanations. The Demographics tab allows users to input parameters (age group, gender, Charlson score, and Guy's stone score). This substratifies the data set and generates a table that displays a percentage probability for each outcome according to the input param-

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

eters, along with the numerator and denominator for each calculation and the overall percentage for each outcome based on the total data set.

The ML Predictions and ML Explanations tabs are interlinked. The ML Predictions tab again allows for input parameters (age, Charlson score, preoperative haemoglobin, Guy's stone score, stone location, size of outer sheath, preoperative midstream urine result, primary puncture site, preoperative DMSA scan, stone size, and image guidance). These factors are fed through each model to generate a single summary table detailing each prediction. This table displays a predicted likelihood as a percentage along with whether the outcome is likely or unlikely (>50% or <50% likelihood).

Within the ML Explanations tab, each model has a separate tab, with an explainer displaying the Shapley weights for each variable. This details and ranks by weight the variables the model is using to predict a particular outcome (Supplementary Fig. 1).

4. Discussion

This is the largest ML study in PCNL to date in terms of the size of the data set and the number of predicted outcomes ($n = 7$), along with temporal validation. We were able to generate an interactive application for data exploration, predictions, and explanatory graphs. The application can be used by clinicians as a decision aid. The models achieved at least moderate accuracy and, for the most part, are well calibrated.

Our study has several limitations, the main one being the heterogeneity of the data set. The BAUS PCNL audit relied on a large number of clinicians entering small amounts of data. This opens the possibility of bias, especially in the reporting of delayed outcomes such as stone-free status. This was reflected in the amount of missing data, which may lead to selection bias. However, it is difficult to generate large data sets of sufficient scale for ML for surgical disciplines. Trial data, the most rigorous and selective data type, are difficult and expensive to scale. Real-world data captured from electronic health records, although much larger in scale, are more variable in their accuracy [30] and therefore the validity is questionable. Audit data lie somewhere between the two, and probably represent the best compromise in terms of data validity and the size of the data set.

The rarity of some outcomes such as visceral injury is another limitation that leads to poorer prediction and calibration of these particular outcomes, despite attempts to deal with this data imbalance (via oversampling and imputation) [31,32], which may lead to overfitting. Addition of imaging data may increase the prediction capability for visceral or vascular injury [33]. Replacement of Guy's stone score [13] with imaging data (eg, CT scans) would probably improve the model performance. Imaging data would capture not only stone complexity (eg, density) but also anatomic complexity (eg, stone location, anatomic variation) [34], which are well-documented factors limiting stone clearance.

The outcome of postoperative complications was also less well calibrated. This is probably because of its status as an aggregate rather than a specific, outcome. However, the model performance on temporal validation was reasonable; external validation is required to check for overfitting.

Stone-free status was poorly defined and relied on clinician definitions. The gold standard for ascertainment of stone-free status is CT imaging [35]. In the UK, patients are often followed up with ultrasound or X-ray scans, which overestimate stone-free status [36]. The definition of stone-free status is also unclear. Historically, fragments of <4 mm or <2 mm were deemed acceptable and included in the definition of stone-free status. However, this has been challenged, as residual fragments can become clinically significant [37] and therefore stone-free status has been redefined as "no fragments" [2]. This ambiguity about stone-free status in terms of ascertainment and the definition used may explain why this particular model performs less well on temporal validation.

Owing to the nature of the data collected (size only rather than multidimensional measures of stone burden) we were unable to conduct comparisons to other nomograms for predicting PCNL outcomes, such as CROES [38] (stone-free status) and STONE [39] (stone-free status and perioperative complications). However, previous studies with smaller data sets have shown the superiority of ML models over these nomograms [6,40].

Future studies should compare the outcomes of our tool to existing nomograms and externally validate the models. Studies aimed at building ML tools for prediction of PCNL outcomes are likely to benefit from the inclusion of imaging data. Integration of other treatment modalities (eg, ureteroscopy) in the online application will enhance its decision aid abilities.

5. Conclusions

This ML study provides the first decision aid tool for seven different PCNL outcomes (https://endourology.shinyapps.io/PCNL_Demographics/). The models are well calibrated with at least moderate accuracy on temporal validation. As the models are temporally valid, they can be used in clinical practice, although further external validation work is needed.

Author contributions: Robert M. Geraghty had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Geraghty, Somani.
Acquisition of data: Finch, Fowler, Rogers, Sriprasad, Smith, Dickinson, Gall.
Analysis and interpretation of data: Geraghty, Thakur, Howles, Somani.
Drafting of the manuscript: Geraghty.
Critical revision of the manuscript for important intellectual content: Geraghty, Thakar, Howles, Finch, Smith, Fowler, Sriprasad, Dickinson, Gall, Rogers, Somani.
Statistical analysis: Geraghty, Thakur.
Obtaining funding: Geraghty.
Administrative, technical, or material support: None.
Supervision: Somani, Thakur, Howles.
Other: None.

Financial disclosures: Robert M. Geraghty certifies that all conflicts of interest, including specific financial interests and relationships and affili-

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011

ations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: This work was supported by the National Institute for Health Research via an academic clinical fellowship and by the Royal College of Surgeons of England via a research fellowship awarded to Robert M. Geraghty. Dr Howles is funded by the Wellcome Trust as a Wellcome Clinical Career Development Fellow. The sponsors played no direct role in the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.euf.2024.01.011.

References

[1] Geraghty RM, Cook P, Walker V, Somani BK. Evaluation of the economic burden of kidney stone disease in the UK: a retrospective cohort study with a mean follow-up of 19 years. BJU Int 2020;125:586–94. https://doi.org/10.1111/bju.14991.

[2] Geraghty RM, Davis NF, Tzelves L, et al. Best practice in interventional management of urolithiasis: an update from the European Association of Urology Guidelines Panel for Urolithiasis 2022. Eur Urol Focus 2022;9:199–208. https://doi.org/10.1016/j.euf.2022.06.014.

[3] Labate G, Modi P, Timoney A, et al. The percutaneous nephrolithotomy global study: classification of complications. J Endourol 2011;25:1275–80. https://doi.org/10.1089/end.2011.0067.

[4] Biswas K, Gupta SK, Tak GR, Ganpule AP, Sabnis RB, Desai MR. Comparison of STONE score, Guy's stone score and Clinical Research Office of the Endourological Society (CROES) score as predictive tools for percutaneous nephrolithotomy outcome: a prospective study. BJU Int 2020;126:494–501. https://doi.org/10.1111/bju.15130.

[5] Shabaniyan T, Parsaei H, Aminsharifi A, et al. An artificial intelligence-based clinical decision support system for large kidney stone treatment. Australas Phys Eng S 2019;42:771–9. https://doi.org/10.1007/s13246-019-00780-3.

[6] Aminsharifi A, Irani D, Tayebi S, Kafash TJ, Shabanian T, Parsaei H. Predicting the postoperative outcome of percutaneous nephrolithotomy with machine learning system: software validation and comparative analysis with Guy's stone score and the CROES nomogram. J Endourol 2020;34:692–9. https://doi.org/10.1089/end.2019.0475.

[7] Aminsharifi A, Irani D, Pooyesh S, et al. Artificial neural network system to predict the postoperative outcome of percutaneous nephrolithotomy. J Endourol 2017;31:461–7. https://doi.org/10.1089/end.2016.0791.

[8] Shapley L. A value for n-person games. Contributions to the theory of games II (1953) 307–317. In: Kuhn HW, editor. Classics in game theory. Princeton, NJ: Princeton University Press; 1997. p. 69–79. 10.1515/9781400829156-012.

[9] Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691. https://doi.org/10.1136/heartjnl-2011-301247.

[10] Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Eur Urol 2015;67:1142–51. 10.1016/j.eururo.2014.11.025.

[11] Armitage JN, Irving SO, Burgess NA. British Association of Urological Surgeons Endourology Section. Percutaneous nephrolithotomy in the United Kingdom: results of a prospective data registry. Eur Urol 2012;61:1188–93. https://doi.org/10.1016/j.eururo.2012.01.003.

[12] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40:373–83. https://doi.org/10.1016/0021-9681(87)90171-8.

[13] Thomas K, Smith NC, Hegarty N, Glass JM. The Guy's stone score—grading the complexity of percutaneous nephrolithotomy procedures. Urology 2011;78:277–81. https://doi.org/10.1016/j.urology.2010.12.026.

[14] Dindo D, Demartines N, Clavien PA. Classification of surgical complications. Ann Surg 2004;240:205–13. https://doi.org/10.1097/01.sla.0000133083.54934.ae.

[15] Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. https://doi.org/10.1136/bmj.m441.

[16] van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019;28:2455–74. https://doi.org/10.1177/0962280218784726.

[17] van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45. https://doi.org/10.18637/jss.v045.i03.

[18] Lunardon N, Menardi G, Torelli N. R package "ROSE": random over-sampling examples. https://rdrr.io/cran/ROSE/man/ROSE-package.html.

[19] Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011.

[20] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.

[21] Krishnapuram B, Shah M, Smola A, et al. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016. ACM Digital Library; 2016. p. 785–94. 10.1145/2939672.2939785.

[22] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65:386–408. https://doi.org/10.1037/h0042519.

[23] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[24] R Core Team. R: a language and environment for statistical computing. https://www.R-project.org/.

[25] Kuhn M. caret: classification and regression training 2021. https://CRAN.R-project.org/package=caret.

[26] Arnold TB. kerasR: R interface to the Keras deep learning library. J Open Source Softw 2017;2. https://doi.org/10.21105/joss.00296.

[27] John CR. MLeval: machine learning model evaluation 2020. https://CRAN.R-project.org/package=MLeval.

[28] Ginestet C. ggplot2: elegant graphics for data analysis. J R Stat Soc Ser A 2011;174:245–6. https://doi.org/10.1111/j.1467-985x.2010.00676_9.x.

[29] Chang W, Cheng J, Allaire J, et al. shiny: web application framework for R 2021. https://CRAN.R-project.org/package=shiny.

[30] Hernandez-Boussard T, Monda KL, Crespo BC, Riskin D. Real world evidence in cardiovascular medicine: assuring data validity in electronic health record-based studies. J Am Med Inform Assoc 2019;26:1189–94. https://doi.org/10.1093/jamia/ocz119.

[31] Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med 2017;376:2507–9. https://doi.org/10.1056/nejmp1702071.

[32] Rich AS, Gureckis TM. Lessons for artificial intelligence from the study of natural stupidity. Nat Mach Intell 2019;1:174–80. https://doi.org/10.1038/s42256-019-0038-z.

[33] Yang G, Wang C, Yang J, et al. Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images. BMC Med Imaging 2020;20:37. https://doi.org/10.1186/s12880-020-00435-w.

[34] Scoffone CM, Cracco CM. Anatomy of the Kidney with Respect to Percutaneous Nephrolithotomy. In: Agrawal MS, Mishra DK, Somani B, editors. Minimally invasive percutaneous nephrolithotomy. Singapore: Springer; 2022. p. 3–15. https://doi.org/10.1007/978-981-16-6001-6_1.

[35] Rob S, Bryant T, Wilson I, Somani BK. Ultra-low-dose, low-dose, and standard-dose CT of the kidney, ureters, and bladder: is there a difference? Results from a systematic review of the literature. Clin Radiol 2017;72:11–5. https://doi.org/10.1016/j.crad.2016.10.005.

[36] Brisbane W, Bailey MR, Sorensen MD. An overview of kidney stone imaging techniques. Nat Rev Urol 2016;13:654–62. https://doi.org/10.1038/nrurol.2016.154.

[37] Brain E, Geraghty RM, Lovegrove CE, Yang B, Somani BK. Natural history of post-treatment kidney stone fragments: a systematic review and meta-analysis. J Urol 2021;206:526–38. https://doi.org/10.1097/ju.0000000000001836.

[38] Smith A, Averch TD, Shahrour K, et al. A nephrolithometric nomogram to predict treatment success of percutaneous nephrolithotomy. J Urol 2013;190:149–56. https://doi.org/10.1016/j.juro.2013.01.047.

[39] Okhunov Z, Friedlander JI, George AK, et al. S.T.O.N.E. nephrolithometry: novel surgical classification system for kidney calculi. Urology 2013;81:1154–60. https://doi.org/10.1016/j.urology.2012.10.083.

[40] Zhao H, Li W, Li J, Li L, Wang H, Guo J. Predicting the stone-free status of percutaneous nephrolithotomy with the machine learning system: comparative analysis with Guy's stone score and the S.T.O.N.E score system. Front Pharmacol 2022;9:880291. https://doi.org/10.3389/fmolb.2022.880291.

Please cite this article as: R.M. Geraghty, A. Thakur, S. Howles et al., Use of Temporally Validated Machine Learning Models To Predict Outcomes of Percutaneous Nephrolithotomy Using Data from the British Association of Urological Surgeons Percutaneous Nephrolithotomy Audit, Eur Urol Focus (2024), https://doi.org/10.1016/j.euf.2024.01.011