*BJU Int* 2024; **133:** 79–86 doi:10.1111/bju.16159



BJUI
BJU International

Original Article

# Predicting obstruction risk using common ultrasonography parameters in paediatric hydronephrosis with machine learning

Adree Khondker[1,2], Jethro C. C. Kwong[3,4], Margarita Chancy[2], Neeta D'Souza[6,7], Kellie Kim[1], Jin K. Kim[2,3], Lai Nam Tse[2], Michael Chua[2,3], Priyank Yadav[2], Lauren Erdman[2,5], John Weaver[6,7], Armando J. Lorenzo[2,3] and Mandy Rickard[2]

*[1]Temerty Faculty of Medicine, University of Toronto, [2]Division of Urology, Department of Surgery, The Hospital for Sick Children, [3]Division of Urology, Department of Surgery, [4]Temerty Centre for AI Research and Education in Medicine, University of Toronto, [5]Vector Institute, Toronto, Onterio, Canada, [6]Department of Urology, Rainbow Babies and Children's Hospital, Case Western Reserve University School of Medicine, Cleveland, OH, and [7]Division of Urology, Children's Hospital of Philadelphia, Philadelphia, PA, USA*

A.K. and J.C.C.K. contributed equally to this work

## Objective

To sensitively predict the risk of renal obstruction on diuretic renography using routine reported ultrasonography (US) findings, coupled with machine learning approaches, and determine safe criteria for deferral of diuretic renography.

## Patients and Methods

Patients from two institutions with isolated hydronephrosis who underwent a diuretic renogram within 3 months following renal US were included. Age, sex, and routinely reported US findings (laterality, kidney length, anteroposterior diameter, Society for Fetal Urology [SFU] grade) were abstracted. The drainage half-times were collected from renography and stratified as low risk (<20 min, primary outcome), intermediate risk (20–60 min), and high risk of obstruction (>60 min). A random Forest model was trained to classify obstruction risk, here named the 'Artificial intelligence Evaluation of Renogram Obstruction' (AERO). Model performance was determined by measuring area under the receiver-operating-characteristic curve (AUROC) and decision curve analysis.

## Results

A total of 304 patients met the inclusion criteria, with a median (interquartile range) age of diuretic renogram at 4 (2–7) months. Of all patients, 48 (16%) were low risk, 102 (33%) were intermediate risk, 156 (51%) were high risk of obstruction based on diuretic renogram. The AERO achieved a binary AUROC of 0.84, multi-class AUROC of 0.74 that was superior to the SFU grade, and external validation ($n$ = 64) binary AUROC of 0.76. The most important features for prediction included age, anteroposterior diameter, and SFU grade. We deployed our application in an easy-to-use application (https://sickkidsurology.shinyapps.io/AERO/). At a threshold probability of 30%, the AERO would allow 66 more patients per 1000 to safely avoid a renogram without missing significant obstruction compared to a strategy in which a renogram is routinely performed for SFU Grade ≥3.

## Conclusions

Coupled with machine learning, routine US findings can improve the criteria to determine in which children with isolated hydronephrosis a diuretic renogram can be safely avoided. Further optimisation and validation are required prior to implementation into clinical practice.

## Keywords

hydronephrosis, PUJ obstruction, nuclear medicine, diuretic renogram, machine learning, artificial intelligence

## Introduction

Diuretic nuclear renography is an accepted method for evaluating suspected PUJ obstruction in children [1–3]. However, this test is an invasive and expensive nuclear medicine scan that is associated with additional costs, radiation exposure, catheterisation and discomfort, in comparison with ultrasonography (US) [3–5]. Moreover, the diagnosis of obstruction may require serial imaging studies with highly standardised protocols [4,6]. For this reason, there has been increased effort to maximise the information obtained from US imaging in the management of hydronephrosis to identify individuals who would benefit most from diuretic renography and those who can be safely monitored with US alone [7–10]. The aim for personalised care and optimisation of US features lends itself naturally to predictive modelling and machine learning (ML) in the context of hydronephrosis [8].

Artificial intelligence (AI) and ML tools have been gaining popularity in paediatric urology and have the potential to influence clinical decision-making [11]. Cerrolaza et al. [12] developed a ML model that used image-based US features to predict half-time ($T_{1/2}$) washout thresholds for obstruction on diuretic renogram with excellent accuracy. Other ML models have predicted PUJ obstruction from renograms, the probability of progression to pyeloplasty, and the risk of recurrent obstruction after pyeloplasty [7,8,13]. While these models have excellent performance, they each require large number of clinical variables, or image pre-processing to generate a prediction.

Here, we aimed to use routine findings from US reports to predict obstruction by diuretic renography thresholds, in effort to select which children may safely avoid a diuretic renogram. With the use of ML, the aim was to develop a model with high sensitivity for obstruction without the need for image processing or technical expertise, and for it to widely accessible by extracting data readily available to providers.

## Patients and Methods

### Study Design and Problem

This study was conducted in accordance with the Standardized Reporting of Machine Learning Applications in Urology (STREAM-URO) framework [14]. Additional details regarding the workflow and analysis are provided in Table S1. This is a supervised multi-class classification problem to predict $T_{1/2}$ drainage times of <20 min, between 20 and 60 min, and >60 min from basic clinical information and routinely reported US findings.

### Data Sources and Eligibility

Following Research Ethics Board approval (REB #1000053438), patients with suspected PUJ obstruction in the context of isolated hydronephrosis were extracted between August 2005 and August 2022 at The Hospital for Sick Children (SickKids, Toronto, Canada) and between 2009 and 2022 from Children's Hospital of Philadelphia (CHOP, Philadelphia, PA, USA). Patients were included if they had isolated hydronephrosis, underwent a diuretic renogram with MAG3 radiotracer and had an US within 3 months prior. Patients were excluded if they had hydroureteronephrosis or concurrent renal anomalies, did not undergo a MAG3 diuretic renogram, had unavailable dates of imaging studies, or had a renogram >3 months after US. As there is a risk of other renal insults or further deterioration above 3 months, we used this as the definition for 'recent' test. The patients from the SickKids dataset were retrieved from a prenatal hydronephrosis database, while the patients from CHOP were retrieved from a suspected PUJ obstruction database. The SickKids dataset was divided into a training and internal validation dataset using a random 85:15 split. An external validation dataset was extracted from patients at CHOP, and patients were included based on identical criteria.

### Feature Generation

Features were abstracted based on information that would be known prior to renogram. Clinical features included age at US and sex. Data from the most recent US prior to renogram included hydronephrosis laterality, anteroposterior diameter (APD), kidney length, and the Society for Fetal Urology (SFU) grade. From the diuretic renogram, the $T_{1/2}$ drainage time was recorded. The final dataset did not contain any missing data. No specific feature engineering or removal of features was performed. Additional information regarding all abstracted data is provided in Table S2.

The primary outcome in this study and ML model was obstruction, defined by $T_{1/2}$ drainage time class: unlikely obstructed (<20 min), prolonged drainage time (20–60 min; considered intermediate-risk category), or obstructed (>60 min; considered high-risk category). As the model objective aims to predict necessity for diuretic renogram, the primary goal was to determine $T_{1/2}$ <20 min, as patients predicted to have drainage >20 min should undergo diuretic renogram. Undergoing pyeloplasty was considered a secondary outcome and was also collected to assess association between $T_{1/2}$ and the clinical decision to proceed with surgical correction. Indications for pyeloplasty at SickKids and CHOP are worsening hydronephrosis (upstaging SFU grade or increasing APD), differential function <40% at baseline, prolonged radiotracer drainage time, and development of symptoms such as infection or pain. Among the CHOP dataset, patients with $T_{1/2}$ drainage >35 min were labelled as prolonged drainage and therefore could not contribute to a high risk (>60 min) category from the SickKids dataset, and only the <20 min criteria was considered.

### Sample Size Calculation

A minimum of 171 patients with 65 demonstrating prolonged drainage time ($T_{1/2}$ >20 min) are required to achieve 80% power to detect a difference of 0.1 in area under the receiver-operating-characteristic curve (AUROC), assuming at least a 38% incidence of obstruction found in the study by Cerrolaza et al. [12].

### Model Selection, Interpretation, and Evaluation

A random Forest classifier ('scikit-learn', version 1.1.3) was trained on our dataset of clinical and US features to predict $T_{1/2}$ drainage time using Python (version 3.8) [15]. This method takes a majority vote using multiple decision trees to predict the outcome. Five-fold stratified cross-validation was used for model training and hyperparameter tuning on the training set, using AUROC as the optimisation metric. Additional information on hyperparameter tuning is provided in Table S3. Model performance was determined by AUROC and decision curves analysis (AUROC is a measure of discriminative performance commonly used in predictive modelling, whereby 0.5 represents random guessing while 1.0 represents perfect discrimination). For comparison, the ML model was compared against SFU grade and APD independently, which has been used as the reference standard in other studies [12]. Bias assessment was determined by comparing AUROC across clinically relevant subgroups including age group (<3 vs ≥3 months), sex, kidney laterality, and SFU grade. All confidence intervals and *P* values were determined using 1000 bootstrap samples with replacement. SHapley Additive exPlanations (SHAP) were used for model explainability. Basic descriptive statistics were tabulated with R (R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org/). Median and interquartile range (IQR) was preferred for reporting of continuous outcomes. Fisher's exact test was used for comparison of dichotomous outcomes. The ML model is referred to as the 'AI Evaluation of Renogram Obstruction' (AERO), which was deployed with R Shiny (R version 4.0.5).

## Results

### Patient Characteristics

A total of 692 children with isolated hydronephrosis were identified. Of these, 344 children had at least one nuclear renogram. Subsequently, 15 and 25 children were excluded for having diethylenetriaminepentaacetic acid (DTPA)-tracer scan and unavailable US within 3 months prior to renogram, respectively. The final dataset included 304 children (SickKids, 258 for training, 46 for internal validation) and 64 children (CHOP) for external validation. Baseline characteristics and radiographic measures are provided in Table 1.

**Table 1** Baseline characteristics and measures for the study population.

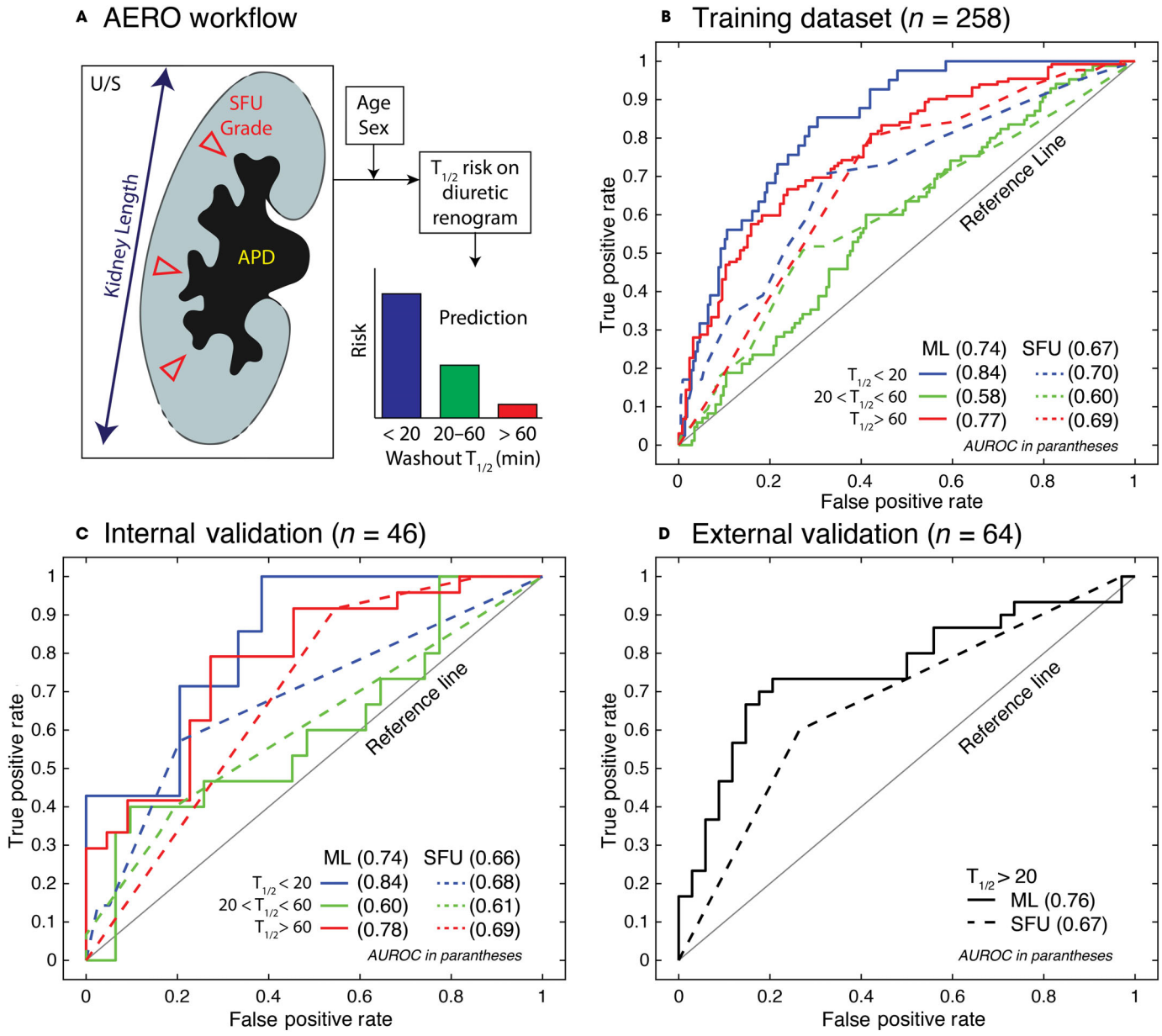
| Variable | SickKids | CHOP |
|---|---|---|
| *N* | 304 | 64 |
| Age at baseline, months, median (IQR) | 2 (1–5) | – |
| Female gender, *n* (%) | 71 (23) | 12 (19) |
| **US findings** | | |
| Age at US, months, median (IQR) | 3 (1–6) | 1 (1–4) |
| Laterality of worse-sided hydronephrosis, *n* (% left) | 194 (64) | 40 (63) |
| Kidney length, mm, median (IQR) | 69 (63–78) | 63 (59–69) |
| APD, mm, median (IQR) | 17 (12–24) | 16 (12–21) |
| Hydronephrosis severity (SFU), *n* (%) | | |
| Grade 1–2 | 26 (9) | 1 (2) |
| Grade 3 | 85 (28) | 36 (56) |
| Grade 4 | 193 (63) | 27 (42) |
| **Diuretic renography** | | |
| Age at diuretic renogram, months, median (IQR) | 4 (2–7) | 3 (2–5) |
| Obstruction definition, $T_{1/2}$ threshold, *n* (%) | | |
| Unlikely, <20 min | 48 (16) | 35 (55) |
| Intermediate risk, 20–60 min | 100 (33) | 29 (45) |
| High risk, >60 min | 156 (51) | 0 (0) |
| **Pyeloplasty** | | |
| Number undergoing pyeloplasty (%) | 181 (59) | 9 (14) |
| Number undergoing surgery within 3 months of renogram (%) | 159 (52) | 0 (0) |
| Age at pyeloplasty for those undergoing pyeloplasty, months, median (IQR) | 6 (4–12) | 17 (9–28) |

*No missing data in either dataset.*

The median (IQR) age at baseline visit was 2 (1–5) months, with an overall follow-up of 35 (19–57) months. US that met inclusion criteria for model training occurred at a median (IQR) age of 3 (1–6) months, and median (IQR) age at diuretic renogram was 4 (2, 7) months. At time of US, the SFU grades in the training cohort consisted of 26 (9%) SFU Grade 2, 85 (28%) SFU Grade 3, and 193 (63%) SFU Grade 4. Of these renograms, the $T_{1/2}$ was >20 min for 256 (84%) patients. A sample of 64 patients from an external institution were retrieved (Table 1) with a significantly larger proportion of patients with $T_{1/2}$ drainage <20 min ($P$ <0.01).

Of 304 patients (SickKids), 186 (61%) underwent pyeloplasty. Among these, 159 (52%) patients underwent pyeloplasty within 3 months of diuretic renogram. The median (IQR) age at pyeloplasty was 6 (4–12) months. Of the patients undergoing a pyeloplasty within 3 months, 152 (96%) had a documented $T_{1/2}$ >20 min. Patients undergoing pyeloplasty had greater proportion of patients with $T_{1/2}$ >20 min (95% vs 66%, $P$ <0.001) and proportion with elevated differential

**Fig. 1** (**A**) Workflow for the AERO, (**B**) ROC curves for internal validation on five-fold cross-validation, (**C**) holdout validation, and (**D**) external validation cohort.



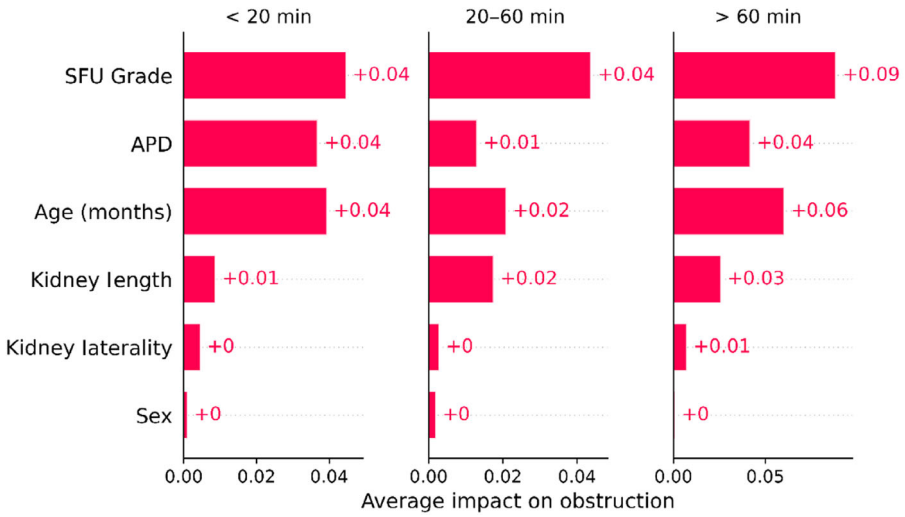

renal function (55% vs 39%, $P = 0.006$), as shown in Table S4.

## Machine Learning Model Performance

The AERO (Fig. 1A) achieved an AUROC of 0.74 (95% CI 0.69–0.77) and 0.74 (95% CI 0.63–0.85) in the training (five-fold stratified cross-validation) and internal validation datasets, respectively (Fig. 1B,C). The AERO outperformed SFU Grade 3/4 and APD >15 mm independently, which achieved an AUROC of 0.67 (95% CI 0.61–0.72, $P$ <0.01) and 0.68 (95% CI 0.62–0.73, $P = 0.01$) for the training datasets, respectively (Table S5). From SHAP explainability, the most important features for the AERO were age, APD, and SFU grade (Fig. 2). No patients within the external validation dataset had $T_{1/2}$ >60 min, therefore the AERO predicted risk of unlikely ($T_{1/2}$ <20 min) vs elevated risk ($T_{1/2}$ >20 min). Among the external dataset, the AERO achieved an AUROC of 0.76 (95% CI 0.63–0.88) compared to 0.67 (95% CI 0.56–0.79, $P = 0.03$) for SFU grade (Fig. 1D).

To contextualise the potential clinical benefits of the AERO, decision curve analysis was conducted to compare the number of renograms avoided if a universal policy was

**Fig. 2** Feature importance rankings highlighting the average impact on probability of obstruction, based on mean SHAP value. SHAP improves understanding of individual predictions by fitting a unique linear model to the AERO and determining corresponding feature contributions.




adopted whereby all patients predicted to have a high-risk probability of obstruction ($T_{1/2}$ >60 min) of ≥30% will undergo a renogram. At this threshold, the AERO would allow 103, 66, and 62 more patients per 1000 to safely avoid a renogram without missing any scans with $T_{1/2}$ >60 min, compared to a strategy for performing renograms in all patients, SFU Grade ≥3, or APD >15 mm, respectively (Fig. 3). Among the training dataset ($n$ = 258), our model predicts that 20 (7.8%) patients could safely defer a diuretic renogram (i.e., have a predicted $T_{1/2}$ <20 min). Of these 20 patients, one (5%) had $T_{1/2}$ >60 min, which would be a critical error. There were no significant differences in the AERO performance when stratified by age group, sex, kidney laterality and SFU grade (Table S6).

We deployed our application in an easy-to-use web- and mobile-application (https://sickkidsurology.shinyapps.io/AERO/). The complete model classification trees are publicly available through the web application.

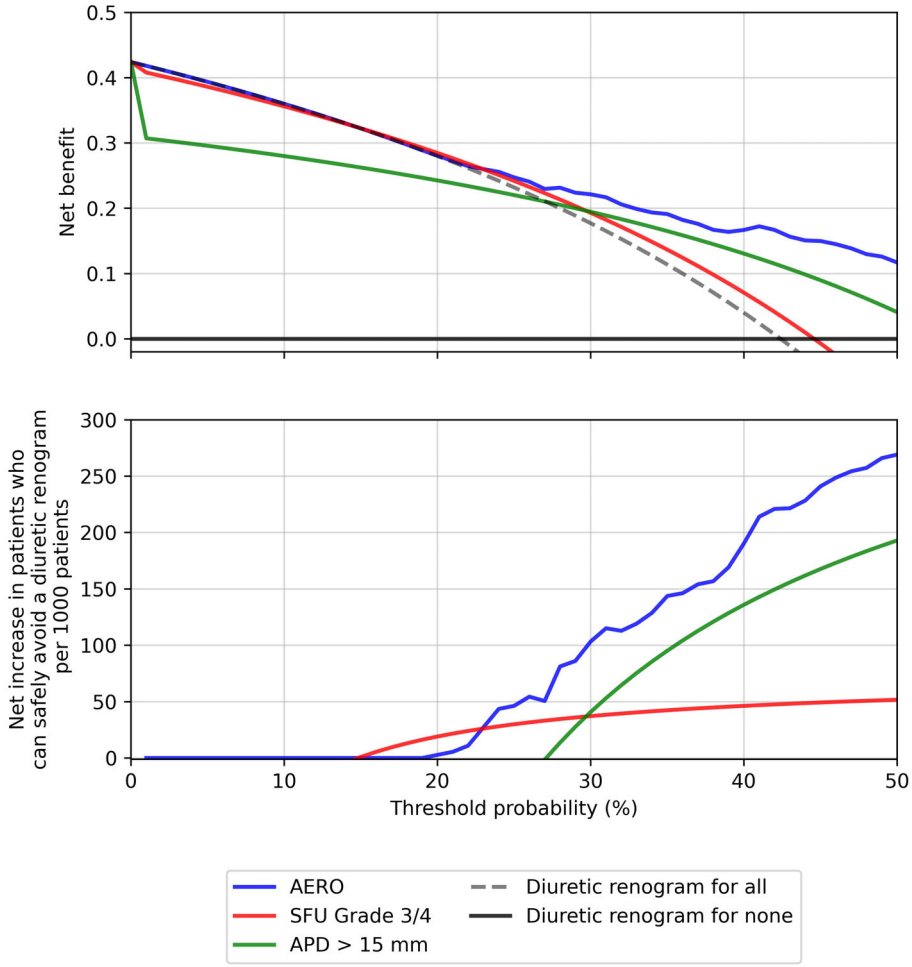
## Discussion

In the present study, we describe a ML model, AERO, which uses routine US information to determine whether a diuretic renogram can be safely avoided. By using a sensitive threshold of T½ clearance of >20 min, the AERO achieved a reasonable performance. Most importantly, the AERO provides modest clinical benefit compared to a strategy where a renogram is performed for patients with SFU Grade ≥3, which translates into 66 renograms that could be safely avoided per 1000 patients. The AERO would be most useful for children with persistent isolated hydronephrosis who are likely to be investigated with a diuretic renogram despite a high likelihood of interval resolution. Furthermore, the AERO was fair across clinically relevant patient- and disease-specific subgroups.

A previous model by Cerrolaza et al. (2016) [12], used both transverse and sagittal US images and processed their images into 131 morphological parameters to predict clearance times of 20, 30, and 40 min on renogram. Their model demonstrated excellent performance with an AUROC of 0.96 but was limited by a small sample size and low model accessibility. While the AERO had a weaker AUROC of 0.76–0.84, it is more accessible due to its simpler input requirements and more accessible as there is no image processing required. Moreover, the AERO performed similarly in an external validation cohort, which supports its generalisability. Of note, no patients in the external validation dataset had $T_{1/2}$ >60 min, which is likely due to differing thresholds for indicating renal scans and surgery. However, even with different institutional practices, our model provides some increased benefit than using SFU Grade ≥3 as an indication for diuretic renogram. This work does suggest that best models in paediatric urology should be trained from data across multiple institutions to support their generalisability.

There is significant controversy on the definition of clinically significant hydronephrosis, obstruction on diuretic renogram, and for indications to perform a pyeloplasty [3,5,16,17]. We do show that 96% of patients who underwent pyeloplasty had $T_{1/2}$ >20 min, and that there is overlap between drainage time on renogram and the decision to perform surgery [18,19]. Other indications for surgery from renograms include significant split differential function or tracer retention, and our model does not predict these outcomes [18,20]. Clinicians should consider additional renogram findings in the decision to perform surgery [21].

**Fig. 3** Decision-curve showing net benefit and net increase in patients who can safely avoid diuretic renogram based on the decision criteria, with the AERO vs SFU Grade 3/4 or APD > 15 mm criteria.




This work must be interpreted in the context of several limitations. First and most importantly, our model was developed on a sample of patients with hydronephrosis who underwent diuretic renogram. This inherently creates a selection bias, as patients with more significant hydronephrosis are more likely to be included in this model and this is evident by the high event rate of elevated drainage time or pyeloplasty within the sample. We attempted to overcome this by using a low threshold of $T_{1/2}$ >20 min with the objective of screening out children who are at lower risk. Next, our sample size of 368 total renograms is larger than previous work on this research problem, but small in the context of ML studies, which limits the generalisability of this model. Despite this, the AERO performed similarly across the training (five-fold stratified cross-validation), internal, and external validation cohorts. Additionally, the AERO uses only five parameters to predict drainage time thresholds and there are additional features that can guide clinical decision-making, such as trend in US findings over time. Within a retrospective dataset, we were unable to incorporate multiple US examinations per patient input due to heterogeneity in the follow-up interval. Lastly, there is subjectivity in the definition of clinically significant hydronephrosis, as described above, which affects the validity of our model. We developed the AERO to predict renogram obstruction by drainage thresholds rather than the subjective yet clinically important decision of performing surgical intervention, and while we show overlap in these outcomes, our model is limited in prediction of this important outcome.

## Conclusion

With the AERO, simple US findings can be used to determine if a diuretic renogram can be safely avoided from the likelihood of obstruction. We demonstrate an association between SFU grade, APD, and age with the likelihood of elevated $T_{1/2}$ drainage. Although this model has lower accuracy than more involved imaging-based models, it offers

moderate benefit with more accessible variables for ML applications. Although further impact testing is warranted, this model can improve clinical decision-making in the management of paediatric hydronephrosis.

## Disclosure of Interests

None.

## Funding

None.

## References

1 **Eskild-Jensen A**, **Gordon I**, **Piepsz A**, **Frøkiaer J**. Congenital unilateral hydronephrosis: a review of the impact of diuretic renography on clinical treatment. *J Urol* 2005; 173: 1471–6

2 **Capolicchio J-P**, **Braga LH**, **Szymanski KM**. Canadian Urological Association/Pediatric Urologists of Canada guideline on the investigation and management of antenatally detected hydronephrosis. *Can Urol Assoc J* 2018; 12: 85–92

3 **Bayne CE**, **Majd M**, **Rushton HG**. Diuresis renography in the evaluation and management of pediatric hydronephrosis: what have we learned? *J Pediatr Urol* 2019; 15: 128–37

4 **Gordon I**, **Piepsz A**, **Sixt R**. Guidelines for standard and diuretic renogram in children. *Eur J Nucl Med Mol Imaging* 2011; 38: 1175–88

5 **Conway JJ**, **Maizels M**. The "well tempered" diuretic renogram: a standard method to examine the asymptomatic neonate with hydronephrosis or hydroureteronephrosis. A report from combined meetings of The Society for Fetal Urology and members of The Pediatric Nuclear Medicine Counc. *J Nucl Med* 1992; 33: 2047–51

6 **Riccabona M**, **Avni FE**, **Blickman JG et al.** Imaging recommendations in paediatric uroradiology: minutes of the ESPR uroradiology task force session on childhood obstructive uropathy, high-grade fetal hydronephrosis, childhood haematuria, and urolithiasis in childhood. ESPR Annual Congress, Edinburgh, UK, June 2008. *Pediatr Radiol* 2009; 39: 891–8

7 **Erdman L**, **Skreta M**, **Rickard M** et al. Predicting obstructive Hydronephrosis based on ultrasound alone. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer Science and Business Media Deutschland GmbH, 2020: 493–503

8 **Lorenzo AJ**, **Rickard M**, **Braga LH**, **Guo Y**, **Oliveria JP**. Predictive analytics and modeling employing machine learning technology: the next step in data sharing, analysis, and individualized counseling explored with a large prospective prenatal hydronephrosis database. *Urology* 2019; 123: 204–9

9 **Shapiro SR**, **Wahl EF**, **Silberstein MJ**, **Steinhardt G**. Hydronephrosis index: a new method to track patients with hydronephrosis quantitatively. *Urology* 2008; 72: 536–8

10 **Cerrolaza JJ**, **Peters CA**, **Martin AD**, **Myers E**, **Safdar N**, **Linguraru MG**. Ultrasound based computer-aided-diagnosis of kidneys for pediatric hydronephrosis. In *Medical Imaging 2014: Computer-Aided Diagnosis*. SPIE, 2014: 733–8

11 **Khondker A**, **Kwong J**, **Malik S et al.** The state of artificial intelligence in pediatric urology: a narrative review. *Front Urol* 2022; 2: 1024662

12 **Cerrolaza JJ**, **Peters CA**, **Martin AD**, **Myers E**, **Safdar N**, **Linguraru MG**. Quantitative ultrasound for measuring obstructive severity in children with hydronephrosis. *J Urol* 2016; 195: 1093–9

13 **Drysdale E**, **Khondker A**, **Kim JK et al.** Personalized application of machine learning algorithms to identify pediatric patients at risk for recurrent ureteropelvic junction obstruction after dismembered pyeloplasty. *World J Urol* 2022; 40: 593–9

14 **Kwong JCC**, **McLoughlin LC**, **Haider M et al.** Standardized reporting of machine learning applications in urology: the STREAM-URO Framework. *Eur Urol Focus* 2021; 7: 672–82

15 **Pedregosa F**, **Varoquaux G**, **Gramfort A et al.** Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30

16 **Eskild-Jensen A**, **Gordon I**, **Piepsz A**, **Frøkiær J**. Interpretation of the renogram: problems and pitfalls in hydronephrosis in children. *BJU Int* 2004; 94: 887–92

17 **Weitz M**, **Portz S**, **Laube GF**, **Meerpohl JJ**, **Bassler D**. Surgery versus non-surgical management for unilateral ureteric-pelvic junction obstruction in newborns and infants less than two years of age. *Cochrane Database Syst Rev* 2016; 7: CD010716

18 **Hodhod A**, **Turpin S**, **Petrella F**, **Jednak R**, **El-Sherbiny M**, **Capolicchio J-P**. Validation of modified diuretic drainage times criteria in congenital hydronephrosis. *J Pediatr Urol* 2021; 17: 832.e1–832.e8

19 **Tabari AK**, **Atqiaee K**, **Mohajerzadeh L et al.** Early pyeloplasty versus conservative management of severe ureteropelvic junction obstruction in asymptomatic infants. *J Pediatr Surg* 2020; 55: 1936–40

20 **Ross SS**, **Kardos S**, **Krill A et al.** Observation of infants with SFU grades 3–4 hydronephrosis: worsening drainage with serial diuresis renography indicates surgical intervention and helps prevent loss of renal function. *J Pediatr Urol* 2011; 7: 266–71

21 **Blum ES**, **Porras AR**, **Biggs E et al.** Early detection of Ureteropelvic junction obstruction using signal analysis and machine learning: a dynamic solution to a dynamic problem. *J Urol* 2018; 199: 847–52

Correspondence: Armando J. Lorenzo, Hospital for Sick Children, 555 University Avenue, Toronto, Onterio M5G 1X8, Canada.

e-mail: armando.lorenzo@sickkids.ca

Abbreviations: AERO, AI Evaluation of Renogram Obstruction; AI, artificial intelligence; APD, anteroposterior diameter; AUROC, area under the receiver-operating-characteristic curve; CHOP, Children's Hospital of Philadelphia; IQR, interquartile range; ML, machine learning; SFU, Society for Fetal Urology; SHAP, SHapley Additive exPlanations; SickKids, The Hospital for Sick Children; STREAM-URO, Standardized Reporting of Machine Learning Applications in Urology; $T_{1/2}$, half-time; US, ultrasonography.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Table S1** The STREAM-URO framework.

**Table S2** Data dictionary for variables used in this work.

**Table S3** Hyperparameter search space used in the development of the AERO, which was built using a random Forest classifier ('scikit-learn' version 1.1.3) on the training dataset. Hyperparameters were selected using 'Optuna', version 2.10.0, with 'TPESampler' and 100 trials (no early stopping). Within each trial, five-fold stratified cross-validation of the

training cohort was performed and the hyperparameters that yielded the highest mean AUROC were recorded. The hyperparameters with the highest mean AUROC across all 100 trials were used in the AERO. Folds were stratified to preserve the proportion of each label class. Hyperparameters with no listed search space values were set to their default values.

**Table S4** Comparison between patients who underwent pyeloplasty vs conservative management, within internal validation dataset.

**Table S5** Model performance and comparisons between the AERO model, SFU grade, and APD over each dataset used in this work.

**Table S6** Variation in the AUROC of the AERO when stratified by age group, sex, kidney laterality, and SFU grade. No statistically significant differences were observed in each subgroup. All confidence intervals and *P* values were determined using 1000 bootstrap samples with replacement.