

Random survival forest models predict renal complications in antenatal hydronephrosis

David A. Ostrowski,^{1,2} Joseph R. Logan,^{3,4,5} Austin Thompson,⁶ Reiley Broms,³ Thomas Syphan,⁶ Maria Antony,³ Mandy Rickard,⁷ Lauren Erdman,^{8,9} Dennis Head,¹⁰ Jessica H. Hannick,⁶ Lynn L. Woo,⁶ Frederick D. Grant,¹¹ Neeta D'Souza,³ Satish E. Viswanath,¹² Chris A. Flask,¹³ Armando J. Lorenzo,⁷ Yong Fan,¹⁴ Gregory E. Tasian,^{2,3,5} John K. Weaver^{6*}

¹Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

²Division of Urology, Department of Surgery, University of Pennsylvania Health System, Philadelphia, PA, USA

³Division of Urology, Department of Surgery, Children's Hospital of Philadelphia, Philadelphia, PA, USA

⁴Translational Research Informatics Group, Children's Hospital of Philadelphia, Philadelphia, PA, USA

⁵Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

⁶Center for Pediatric Urology, Glickman Urological and Kidney Institute, Cleveland Clinic, Cleveland, OH, USA

⁷Division of Urology, Hospital for Sick Children, Toronto, ON, Canada

⁸James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

⁹Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, OH, USA

¹⁰Department of Urology, Northeast Ohio Medical University, Rootstown, OH, USA

¹¹Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA, USA

¹²Department of Pediatrics and Biomedical Engineering, Emory University, Atlanta, GA, USA

¹³Department of Radiology, Case Western Reserve University School of Medicine, Cleveland, OH, USA

¹⁴Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

OSTROWSKI DA, LOGAN JR, THOMPSON A, BROMS R, SYPHAN T, ANTONY M, RICKARD M, ERDMAN L, HEAD D, HANNICK JH, WOO LL, GRANT FD, D'SOUZA N, VISWANATH SE, FLASK CA, LORENZO AJ, FAN Y, TASIAN GE, WEAVER JK. Random survival forest models predict renal complications in antenatal hydronephrosis. *Can J Urol* 2026;33(5):1261–1272.

Objectives: This study aims to build machine learning models to analyze Technetium-99m mercaptoacetyl triglycine (MAG3) renal scans and clinical variables from patients with antenatal hydronephrosis (ANH) with concern for ureteropelvic junction obstruction (UPJO) and equivocal initial MAG3 renal scans to predict the risk of renal complication development.

Methods: A retrospective cohort of patients under 1-year-old with ANH concerning for UPJO with an equivocal MAG3 renal scan evaluated at our institution from January 2009 to June 2021 was identified. Equivocal renal scans were defined as studies where the affected kidney demonstrated split function greater than 39%, and subsequent clinical recommendation was continued surveillance instead of surgery. Random survival forest models were constructed to predict the risk of renal

complications (decreased function, increased parenchymal thinning, worsening hydronephrosis) using clinical variables (Clinical Model), or numerical features extracted from MAG3 renal scans (Imaging Model), or clinical and numeric renal scan features (Ensemble Model).

Results: One-hundred fifty-two patients were included; 62 patients developed a renal complication, and 90 did not develop a renal complication with a minimum of 3-year follow-up. The Imaging Model (C-index 0.74; 95% CI: 0.68–0.79) and the Clinical Model (C-index 0.70; 95% CI: 0.66–0.75) performed similarly, while the Ensemble Model (C-index 0.76; 95% CI: 0.75–0.86) demonstrated improved performance compared to either individually. Differential function per unit volume of the affected kidney was the most important driver of the Clinical Model's predictions. All models identified high-risk patients at significantly higher risk of renal complications on Kaplan-Meier log-rank analysis ($p < 0.0001$).

Conclusion: Random survival forest models using clinical and/or renal scan features may improve the prediction of renal complication risk in patients with ANH, with concern for UPJO and an equivocal initial MAG3 renal scan.

Key Words: diuretic renal scan, antenatal hydronephrosis, machine learning, random survival forest

Received date 05 December 2025

Accepted for publication 31 March 2026

Published online 30 October 2026

*Corresponding Author: John K. Weaver.

Email: jweave2925@gmail.com, weaverj14@ccf.org

Introduction

Antenatal hydronephrosis (ANH) is a common congenital abnormality, affecting 1–5% of all pregnancies.¹ In cases of severe or persistent hydronephrosis on postnatal ultrasound, technetium-99m mercaptoacetyltriglycine (MAG3) diuretic renograms are utilized to evaluate potential urinary tract obstruction, most commonly ureteropelvic junction obstruction (UPJO), that can cause renal damage if untreated.^{2–5} However, a wide range of indeterminate post-diuretic radiotracer washout half-times ($t_{1/2}$) for ANH patients has been reported, due, in part, to infant physiology.^{6–10} Therefore, predicting the risk of renal complications based on diuretic renograms is challenging, particularly when $t_{1/2}$ is indeterminate for obstruction (i.e., 10–40 min).^{6,7,11} Currently, this patient population is currently followed by frequent serial clinical exams, repeat renal scans, and serial renal ultrasounds that are driven by an individual provider's clinical gestalt and with frequency that can be burdensome for patients and their families.¹² Thus, there exists a need for tools to improve the risk assessment of ANH patients with indeterminate renal scans.

Machine learning algorithms have shown promise in pediatric urology,^{13–15} such as conducting time-to-event analysis of chronic kidney disease progression risk for children with posterior urethral valves.^{16,17} Machine learning has been used in ANH to grade hydronephrosis severity,¹⁸ predict clinical progression based on ultrasound,^{19,20} and help improve MAG3 renal scan interpretation by predicting which patients would progress to a renal complication.^{21,22} Prior binary classification models demonstrated potential to improve diuretic renogram prognostication but provided limited insight into an individual's risk or the time course of renal complication development.

Our goal was to develop random survival forest machine learning models to analyze MAG3 renal scans and/or clinical variables from ANH patients with equivocal initial MAG3 diuretic renograms to predict an individual's risk of renal complications. We hypothesized that our machine learning models would predict renal complication risk with strong performance for a currently prognostically challenging patient population and that a model using both renal scan and clinical variables would achieve greater performance than models using either data set alone.

Methods

Study design

A retrospective cohort study was performed on a population of pediatric patients with ANH concerning for UPJO with at least one available MAG3 renal scan evaluated at the Children's Hospital of Philadelphia (CHOP) between January 2009 and June 2021. This study was reviewed and approved by CHOP Institutional Review Board (IRB#: 17-014098). Given the retrospective and de-identified nature of this study, the CHOP IRB approved a waiver of informed consent.

Population

A population of 7591 consecutive patients less than one year of age diagnosed with ANH was identified from our institution's database using International Classification of Disease 9th and 10th Edition diagnostic codes (ICD-9 753.29, ICD-10 Q62.0). Patients were included if they had hydronephrosis on renal ultrasound that was Society of Fetal Urology grade ≥ 3 or Urinary Tract Dilation (UTD) grade ≥ 2 without ureteral dilation and an equivocal renal scan; and either developed a renal complication (defined below) or had 3 years minimum of clinical follow-up without a renal complication. For patients who did not develop a complication, a minimum of 3 years of follow-up was required for study inclusion to ensure that patients with insufficient follow-up in the setting of a retrospective cohort were excluded. Equivocal renal scans were defined as studies in which the affected kidney demonstrated an indeterminate $t_{1/2}$ (>10 min) and split function greater than 39% and the subsequent clinical recommendation was continued surveillance instead of surgery based on the holistic interpretation of the patient's imaging, laboratory results, and overall clinical condition by the fellowship-trained pediatric urologist evaluating the patient at that point in time. Patients with insufficient follow-up, history of known or suspected ipsilateral vesicoureteral reflux, bilateral UPJO, or renal anatomic abnormalities, such as duplicated systems, renal malrotation, renal ectopia, or horseshoe kidney, were excluded. Patients were followed with a mix of ultrasound and MAG3 renal scans, with intervals determined by the judgment of the attending pediatric urologist at the time of each clinical evaluation of the patients in this retrospective cohort.

Exposure

The exposure was each patient's earliest equivocal MAG3 renal scan, including numerical features extracted by the machine learning algorithm from the

renal scan and relevant clinical variables available at that time.

Outcome

The outcome was renal complications occurring >6 months after an equivocal MAG3 renal scan, including: >5% decline in renal function of the affected kidney on follow-up renal scan, follow-up renal ultrasound showing increased parenchymal thinning, worsening hydronephrosis (defined as an increase UTD hydronephrosis classification per radiologist review), new onset flank pain (based on the clinician's physical exam or patient report for patients able to verbalize symptoms), or development of pyelonephritis (determined by the evaluation of the attending pediatric urologist at the time considering patient's clinical condition, fever, and urinalysis; there was no standardized method for urine sample collection amongst the cohort). Decline in function >5% on subsequent renal scan was chosen as a threshold for meaningfully decreased renal function, given the reported <5% margin of error of MAG3 renal scans.^{23–25} Ultrasounds read as displaying increased parenchymal thinning or hydronephrosis on radiologist impression were reviewed by a fellowship-trained pediatric urologist (author JKW, blinded to machine learning analysis) for validation in the UTD hydronephrosis grading scheme. This outcome of renal complications was chosen as an objective endpoint, as an individual surgeon's decision to proceed to pyeloplasty is intrinsically subjective. Change in $t_{1/2}$ was not included in the definition of a renal complication as there is a wide range of indeterminate $t_{1/2}$ in this population, and so changes in $t_{1/2}$ may not constitute a meaningful renal complication, whereas the outcomes included in our composite endpoint include changes that suggest active or impending meaningful and permanent renal injury that would often prompt a surgeon to proceed with pyeloplasty.

Clinical variables

For eligible patients, relevant clinical characteristics at the time of initial equivocal renal scan were collected from the electronic medical record. Extracted clinical features included age, sex, race, and kidney laterality as well as discrete features from each patient's first equivocal renal scan, including differential function per total volume of the affected kidney, differential function per unit volume of the affected kidney (radiotracer counts normalized by the relative volume of each kidney), and post-diuretic radiotracer count half-life ($t_{1/2}$). The time-to-event was defined as the time from initial MAG3 renal scan

until renal complication development or until the last clinical follow-up if the patient did not develop a complication.

MAG3 renal scan protocol and image processing
MAG3 renal scans were performed at a single center as per the protocol standardized by Nuclear Medicine physicians at our institution using a Philips Forte SPECT Gamma Camera System (Phillip North America Corporation; 222 Jacobs St., Cambridge, MA). Prior to image acquisition, an intravenous line was placed, and bladder catheterization was performed. During MAG3 renal scan image acquisition, images were obtained every second for the first minute after radiotracer infusion and then every 30 s for one hour for a total of 180 images per study. During the scan protocol, normal saline was infused at 15 mL/kg for the first 30 min and then at 5 mL/kg for the final 30 min of the study. One mg/kg furosemide was administered 20 min after radiotracer infusion per our institution's protocol, as described in prior publications from our institution.²² Time activity curves were created using our institution's standard nuclear medicine processing software. Radiotracer counts from each kidney and the bladder were then plotted as a function of time and converted to numerical features (counts from each kidney and bladder as a function of time) using software adapted from parametric MRI (pMRI) developed at our institution.²⁶ The raw features of the renal scans were directly imported into the models by this software without higher-level feature extraction or data manipulation.

Machine learning analysis

Random survival forest models were constructed using Python 3.8.5, scikit-learn 1.0, and scikit-survival 0.17.2 to perform time-to-event analysis for developing renal complications.²⁷ Three random survival forest models were built. The Imaging Model used the numerical features extracted from each patient's initial MAG3 diuretic renogram by pMRI software (radiotracer counts as a function of time as described in the above sub-section titled "MAG3 Renal Scan Protocol and Image Processing"), and the Clinical Model used the extracted clinical variables (as described in the above sub-section entitled "Clinical Variables"). The models were built without a pre-trained algorithm, as no previously reported data using the MAG3 renal scan numeric features extracted by the pMRI software were available. Hyperparameters were tuned via Grid Search of 500 combinations. Utilized hyperparameters included the number of decision trees (final values of 1600 trees for the Imaging Model, 200 trees for the Clinical

Model), the maximum depth of each decision tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be in a leaf node, and the maximum number of features to consider when splitting a node. For building each tree, bootstrapping was set to true. For the Clinical and Imaging models, the study sample was divided into training (80%) and testing (20%) sets, and the results were internally validated with three-fold cross-validation without stratification by event occurrence. The Ensemble Model was generated by performing an equal-weight average of the risk scores of the Imaging and Clinical models, as the vastly different underlying data sets (120 temporally-linked numeric features for the Imaging Model versus 5 independent variables for the Clinical Model) would have limited the utility of including both Clinical and Imaging features in the same random survival forest. Temporal leakage was avoided by using only variables at the time of initial patient presentation as inputs to generate model predictions. In addition, as described in the Results, the Clinical Model and the Imaging Model demonstrated similar performance, and so equal-weighting for development of the Ensemble Model was appropriate. The models' output for each patient consisted of a risk score and a predicted progression-free survival probability function.

Statistical analysis

For clinical variables, dichotomous variables are reported as counts and percentages of the cohort and compared between groups by Fisher's exact test. Continuous variables are reported as median and interquartile range (IQR) and compared between groups by the Mann-Whitney U test.

Predicted progression-free survival probability curves generated by the Ensemble Model were plotted for selected individual patients. To evaluate the time-to-event model's discriminative performance, concordance indices (C-index) were calculated based on cross-validated performance. To evaluate performance accuracy at specific time points, receiver operating curves (ROC) with areas under the curve (AUC) measurements were calculated at 2, 5, and 6 years. For ROC calculation, the models' predictions for the entire cohort at each of the above discrete time points were assessed. Performance metrics of the models are reported with 95% confidence intervals, which were computed via bootstrapping with 1000 iterations. Permutation scoring was used to determine the relative contribution of each clinical variable to the Clinical Model's predictions.

Using the risk scores predicted by each model, a risk score threshold was used to stratify patients into high-risk (top 20%) and low-risk (bottom 80%) groups to demonstrate the efficacy of the model in being able to identify patients at high-risk for renal complications. This threshold was chosen for the purposes of demonstration as it yielded a high specificity for renal complication development while having a robust amount of patients in each risk group. For example, for the Ensemble model, the specificity of the top 20% risk score threshold for identifying patients who would develop a renal complication was 92.1% with a sensitivity of 37%. Proportions of patients developing renal complications in each risk group were compared by the chi-square test. Kaplan-Meier plots and log-rank tests were used to compare progression-free survival between the low-risk and high-risk cohorts for the 149 patients with complete clinical and imaging data.

Statistical analysis was performed with Python 3.8.5 and R 4.2.2 (R packages utilized include "survival", "survminer", "ggplot2", "robustbase", "matrixStats", and "CompareC"). STREAM-URO reporting standards, which detail recommendations for manuscript structure, description of data handling, and presentation of data for investigators using machine learning models, were adhered to in preparation of this manuscript as closely as possible within the limits of the available dataset.²⁸ Software and statistical analysis used in this project are linked in GitHub (https://github.com/loganj2chop/MAG3_surv).

Results

Our final analysis cohort consisted of 152 patients meeting all eligibility criteria. Of the 152 patients, 62 developed a renal complication, and 90 did not, with 3 years minimum follow-up. Three patients did not have all clinical variables available and so were not included in the training and testing sets for the Clinical and Ensemble models. The patients were predominantly male, and the left kidney was most commonly affected (Table 1). Relevant clinical variables and conventional parameters from each patient's initial MAG3 renal scan are reported in Table 1. Patients developing a complication were statistically significantly younger at initial renal scan (2.1 vs. 3.1 months, $p = 0.01$) with longer $t_{1/2}$ (20 vs. 16 min, $p = 0.006$) and decreased differential functions per total volume (48% vs. 50%, $p = 0.02$) and per unit volume (40% vs. 45%, $p < 0.001$) of the affected kidneys though these differences were small and of questionable clinical relevance for clinical

decision-making in a prospective setting (Table 1). Of the patients who developed a renal complication, 16 (26%) had a greater than 5% decline in renal function, 2 (3%) developed pyelonephritis, 2 (3%) developed flank pain, 20 (32%) had worsened parenchymal thinning, and 22 (35%) had worsening hydronephrosis.

All patients who developed a renal complication underwent subsequent pyeloplasty.

All random survival forest models had a C-index greater than 0.7 in predicting the development of renal complications (Table 2). The Ensemble Model (C-index 0.76; 95% CI: 0.75–0.86) demonstrated the

TABLE 1. Clinical characteristics of the overall cohort and of patients who developed a renal complication versus patients who did not develop a renal complication at the time of renal scan, including age at time of renal scan, months, and median (IQR)

Characteristic	All patients (N = 152)	Patients developing a renal complication [#] (N = 62)	Patients not developing a renal complication (N = 90)	p-value
Sex, n (%)				>0.99
Male	116 (76%)	47 (76%)	69 (77%)	
Female	36 (24%)	15 (24%)	21 (23%)	
Affected Kidney Laterality, n (%)				0.48
Left	102 (67%)	44 (70.9%)	58 (64.4%)	
Right	50 (33%)	18 (29.1%)	32 (35.6%)	
UTD Hydronephrosis Grade on Ultrasound Prompting Renal Scan [†] , n (%)				0.4
P2	78 (51%)	35 (56%)	43 (48%)	
P3	73 (48%)	27 (44%)	46 (51%)	
Initial Equivocal MAG3 Renal Scan Findings, median (IQR)				
Age at time of renal scan, months	2.6 (1.8, 6.0)	2.1 (1.6, 5.2)	3.1 (2.1, 7.8)	0.01*
t _{1/2} , minutes	19 (10, 25)	20 (13.25, 30.75)	16 (5, 25)	0.006**
Differential function per total volume of affected kidney, %	49 (47, 52)	48 (44.75, 50.25)	50 (47, 52)	0.02*
Differential function per unit volume of affected kidney, %	43 (39, 47)	40 (35, 45)	45 (41, 47)	<0.001***
Follow-Up				
Follow-up time (time to either develop a renal complication or time to the most recent follow-up), months, median (IQR)	48.2 (28.1, 65.9)	17.1 (12.0, 38.5)	56.2 (48.2, 76.2)	<0.001***
UTD Hydronephrosis grade at time of complication development or at most recent follow-up, n (%)				<0.001***
No dilation	11 (7%)	0	11 (12%)	
P1	29 (19%)	1 (2%)	28 (31%)	
P2	40 (26%)	10 (16%)	30 (33%)	
P3	72 (47%)	51 (82%)	21 (23%)	

Note. Continuous variables are presented as median (IQR). Discrete variables are presented as counts (percentage of cohort). Statistically significant with * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. [#]Breakdown of renal complications: >5% decline in renal function = 16, pyelonephritis = 2, pain = 2, parenchymal thinning = 20, worsening hydronephrosis = 22. [†]One patient who did not develop a renal complication did not have ultrasound images available for review but received a MAG3 renal scan due to clinical concern for urinary tract obstruction.

TABLE 2. Performance metrics of each random survival forest time-to-event model

Model	C-index	AUC 2 years	AUC 5 years	AUC 6 years
Imaging model	0.74 (0.68–0.79)	0.79 (0.74–0.85)	0.77 (0.73–0.83)	0.76 (0.72–0.81)
Clinical model	0.70 (0.66–0.75)	0.77 (0.74–0.84)	0.75 (0.71–0.79)	0.73 (0.70–0.76)
Ensemble model	0.76 (0.75–0.86)	0.82 (0.78–0.86)	0.82 (0.77–0.86)	0.80 (0.75–0.86)

Note. Values in parentheses are the 95% confidence intervals for each performance metric.

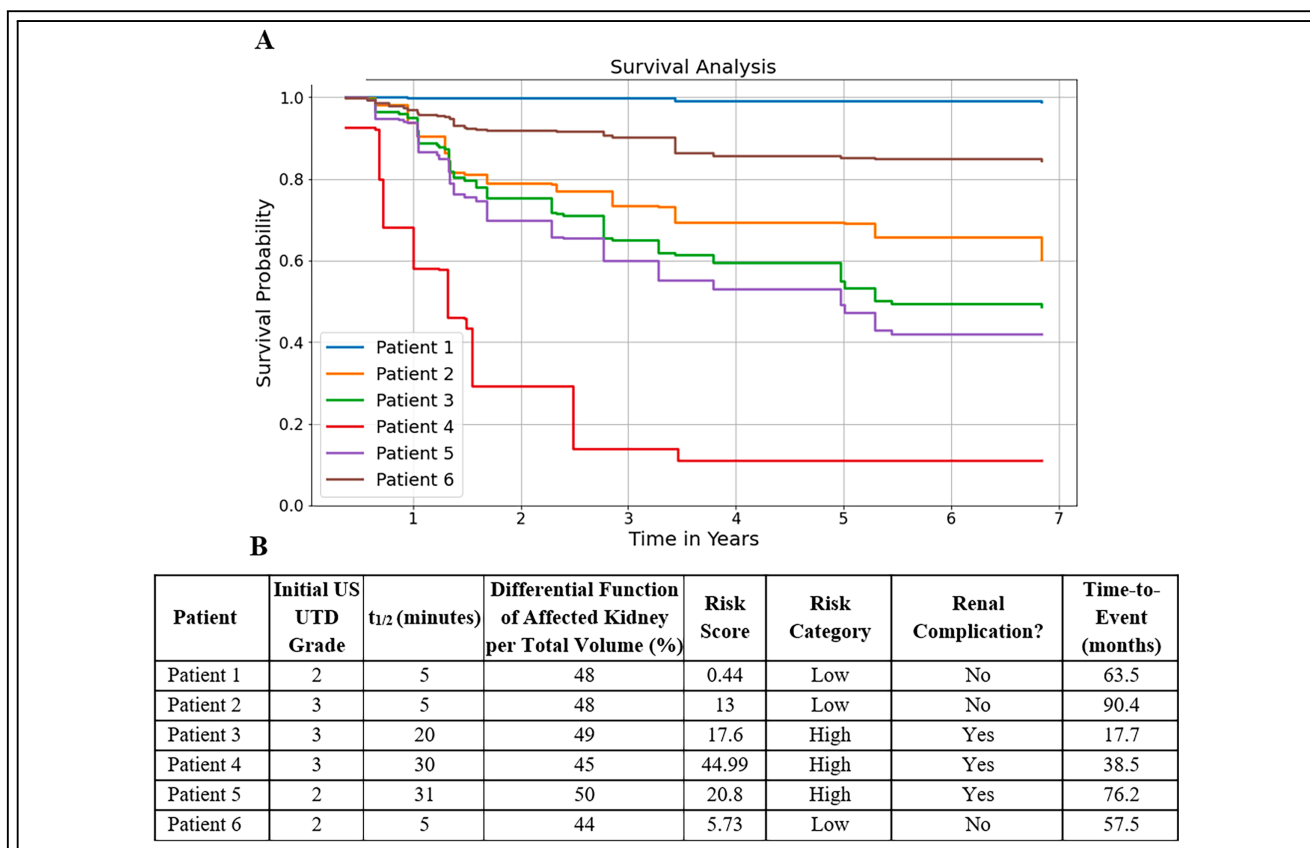


FIGURE 1. Ensemble model results for selected individual patients. (A) Predicted progression-free survival probability curves generated by the Ensemble model for six representative patients with varying risk scores. (B) For each patient included in 1A, the table reports their initial ultrasound UTD hydronephrosis grade, the $t_{1/2}$ and differential function per total volume of the affected kidney on their initial MAG3 diuretic renogram, their risk score, their risk category (high-risk versus low-risk), whether they developed a renal complication, and their time-to-event. Note. For patients experiencing a complication, the time in months from their first equivocal renal scan until complication development; for patients not experiencing a complication, the time in months from their first equivocal renal scan until their last clinical follow-up.

best overall performance, followed by the Imaging Model (C-index 0.74; 95% CI: 0.68–0.79) and the Clinical Model (C-index 0.70; 95% CI: 0.66–0.75) (Table 2). Predicted progression-free survival probability curves output by the Ensemble Model for representative patients are demonstrated in Figure 1.

ROC curves were calculated at 2, 5, and 6 years following each patient's initial equivocal renal scan. The ROC curves at 6 years are demonstrated in Figure A1. AUCs at 2, 5, and 6 years were greater for the Ensemble Model compared to the Clinical Model and similar between the Imaging Model and

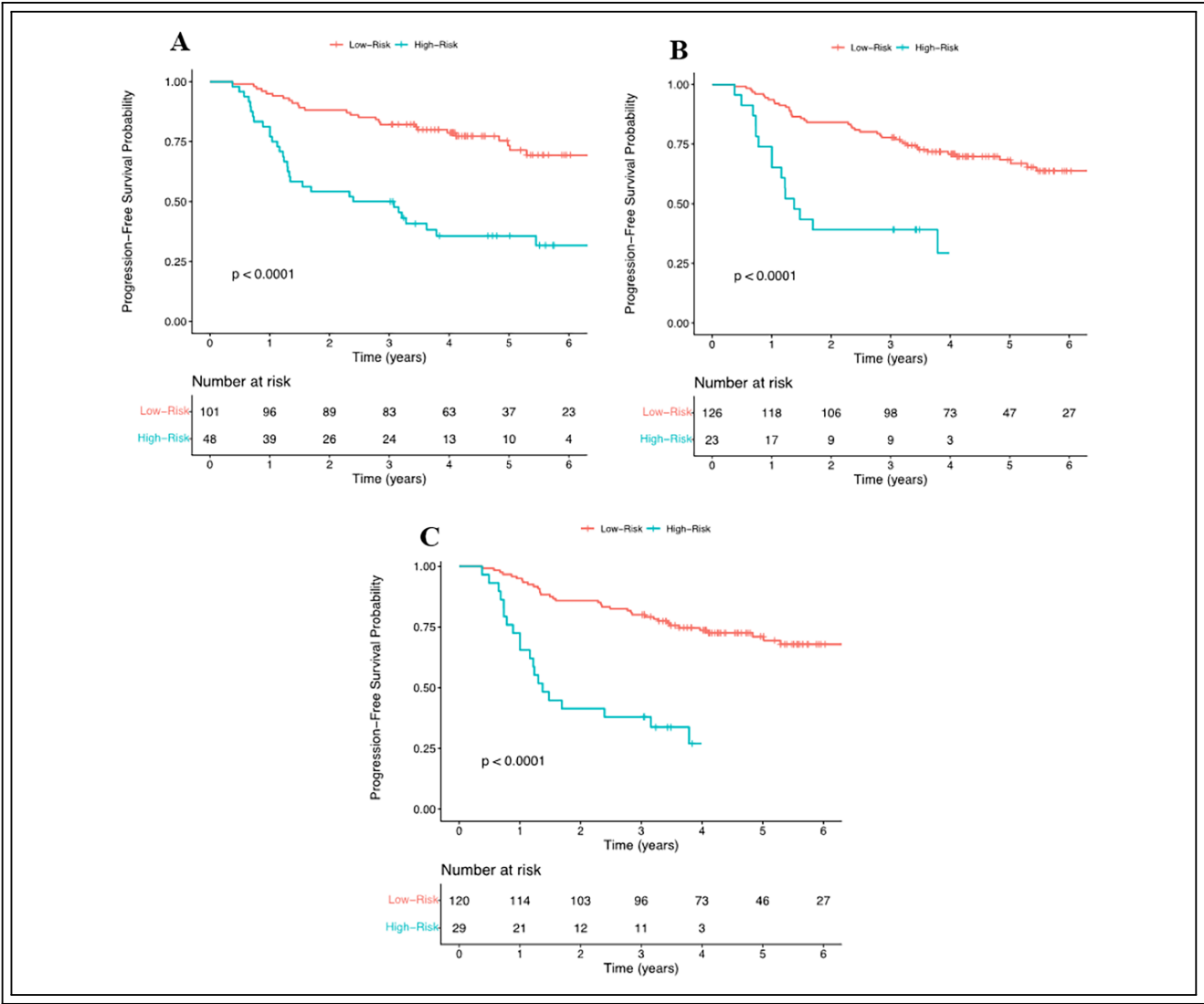


FIGURE 2. Kaplan-Meier plots of progression-free survival probability of patients classified by the machine learning models as low-risk (bottom 80% risk scores) or high-risk (top 20% risk scores) by (A) the Imaging model, (B) the Clinical model, and (C) the Ensemble model. Note. The p -values are the result of log-rank tests comparing progression-free survival between the risk groups for each model.

the Ensemble Model (Table 1). Permutation scoring demonstrated that the differential function per unit volume of the affected kidney was the strongest driver of Clinical Model predictions, followed by $t_{1/2}$ (Table A1).

Using the random survival forest model risk scores to stratify patients as low-risk or high-risk, progression-free survival was decreased for the high-risk groups across all models on Kaplan-Meier and

log-rank analysis (Figure 2; $p < 0.001$). For the Ensemble Model, 22/29 (76%) high-risk versus 38/120 (32%) low-risk patients developed a renal complication ($p < 0.001$). For the Imaging Model, 30/43 (70%) high-risk versus 32/109 (29%) low-risk patients developed a renal complication ($p < 0.001$). For the Clinical Model, 27/40 (67.5%) high-risk versus 33/109 (30.3%) low-risk patients developed a renal complication ($p < 0.001$).

Discussion

Amongst patients with ANH with concern for UPJO with equivocal initial MAG3 renal scans with clinically similar radiotracer washout times and differential function, our machine learning models demonstrated strong performance in risk stratification and individualized time-to-event predictions of renal complication risk. Each of the three models demonstrated moderately high discriminative performance with C-indices greater than 0.7 as well as promising accuracy at multiple time points with AUCs greater than 0.7 at 2, 5, and 6 years. The Ensemble Model demonstrated the strongest C-index performance, followed closely by the Imaging Model and then the Clinical Model, suggesting a benefit of combining clinical variables traditionally assessed by urologists with renal scan features extracted by our algorithm. Risk scoring from all models identified high-risk patients with significantly increased rates of renal complications on Kaplan-Meier analysis. Permutation scoring of Clinical Model variables demonstrated improved predictive value of differential renal function per unit volume compared to traditional split function (differential function per total volume), suggesting a potentially important clinical role for differential function per unit volume when available.

Patients with ANH with concern for UPJO on ultrasonography require further evaluation to determine whether surgery is necessary. MAG3 renal scan is the current gold-standard imaging modality to characterize potential obstruction despite known challenges in interpretation for infants,^{6,11} such as reservoir effects from severely dilated collecting systems,⁸ decreased glomerular filtration rates,^{9,10} and variation due to hydration or bladder drainage status.¹⁰ As a result, current evidence has established a wide range of indeterminate $t_{1/2}$ values, approximately 10–40 min, for patients with ANH, making interpretation and application of renal scan findings difficult in this population.^{6,7,11} Given the central role of renal scans in identifying potential surgical candidates, there is a need for tools to improve diuretic renogram interpretation.

Machine learning is a promising approach to improve the diagnostic and prognostic yield of diuretic renograms. Blum et al. developed a linear support vector machine classifier to analyze diuretic renograms from ANH patients with concern for urinary tract obstruction in order to classify the renograms as from patients with obstruction

requiring surgery versus from patients without obstruction.²¹ In a small retrospective cohort, they reported 93% accuracy for their machine learning model versus 82% maximum accuracy using traditional $t_{1/2}$ and 30-min clearance cutoffs. However, their models required manual selection of drainage curve features and did not include clinical variables. Furthermore, all ANH patients with concern for UPJO were included in the study population, with most patients having $t_{1/2} < 10$ min or > 30 min, making their results less applicable to patients with indeterminate diuretic renograms. Previously, our group identified a cohort of patients with ANH with concern for UPJO with equivocal initial diuretic renal scans and built machine learning models to perform binary classification using a convolutional neural network to analyze diuretic renogram features and a random forest classifier to analyze clinical variables.²² In predicting which patients would develop a renal complication and which would not, the renal scan model achieved 73% binary classification accuracy versus 62% accuracy for the clinical variable model. These studies demonstrated the potential of machine learning to improve diuretic renogram interpretation, though their binary outputs (obstruction versus no obstruction) may oversimplify a complex clinical problem.

Our study built on this foundation by using random survival forest models that perform time-to-event analysis and provide more nuanced risk assessments. Our models empirically extracted numeric features from diuretic renograms without requiring manual selection. They then used these imaging features and/or clinical variables to perform time-to-event analysis of an individual's renal complication risk with strong overall performance. For each patient, the models produced individualized risk scores and predicted progression-free survival probability curves, providing more refined, individualized risk assessment compared to binary classification. By providing an individual risk score and individualized prediction of probability of developing a renal complication, these models can be used to tailor surveillance and follow up plans for an individual patient to help minimize the burden of surveillance of those patients at lower risk of renal complication development and to help with early identification of those individuals at higher risk of renal complications to potentially prompt earlier intervention to prevent complication development. Importantly, our models achieved strong performance among a cohort with

clinically similar traditional diuretic renogram parameters between patients developing versus not developing a renal complication. Thus, the ability of our models to effectively identify high- and low-risk individuals within this cohort represented an improvement compared to traditional MAG3 renal scan interpretation methods and suggested a potential role for our machine learning models as clinical tools to help guide clinical follow-up protocols and to help predict an individual's potential need for and timing of surgery.

Limitations of this study include a retrospective, single-center design and small cohort size, which limit generalizability. However, it is common in the pediatric urology literature, including prior work by our group, to start with a small sample size at a single institution before expanding to multiple-institutional studies once the model has shown promise.^{17,28} Furthermore, our definition of equivocal renal scan does contain some subjectivity by including the clinical decision to continue with surveillance. However, among this patient population in which there is significant uncertainty and subjectivity, this definition is a reasonably objective metric based on clinical practice in a high-volume center of excellence. We included $t_{1/2}$ and differential renal function in our model since they are the most commonly used parameters to evaluate renal obstruction, although other parameters, including upright drainage time, delayed drainage time, and normalized residual activity or output efficiency, have been reported.^{29–31}

As with many machine learning models, there is potential for overfitting and poor generalizability outside of the study cohort, although cross-validation helps mitigate these risks. Our cohort was selected specifically with tight inclusion criteria to evaluate a subset of patients with ANH for whom current methods of risk assessment are lacking, resulting in an overall cohort that, while large by pediatric urology standards, is small by machine learning standards. Thus, our results should not be applied to patients outside of our specific population of interest - children under one year of age with ANH and concern for UPJO with an equivocal initial diuretic renogram. In the future, when we validate our models on larger populations at additional institutions, we will be able to perform sensitivity analyses in which we use stricter outcome parameters (e.g., only patients with a differential function decline of 10%). Also, we plan to pursue further studies using the risk-scoring of our models to help guide follow-up protocols for this

patient population to potentially limit the surveillance burden in patients at low risk for complications and to earlier identify high-risk patients to help prevent renal complications. Other future directions include continued expansion of the model to include clearly obstructed and non-obstructed patients, prediction of individual complications rather than just a composite endpoint, calibration and validation of the models with additional patients, narrowing of findings in composite outcomes to specific outcomes, inclusion of potential multi-institutional collaborations (with the understanding that this would likely involve standardization of renal scan protocols across institutions to ensure adequate model performance), and prospective trials.

Conclusions

We developed random survival forest machine learning models to analyze MAG3 renal scans and/or relevant clinical variables from ANH patients under one year of age with an initial equivocal renal scan to predict an individual's renal complication risk with strong overall performance. These results suggest a potential future role for machine learning models as tools providing individualized risk assessments to help guide management of a subset of ANH patients for whom current prognostic tools provide limited information.

Acknowledgement

The authors would like to acknowledge Dmitry Khrichenko for his technical assistance with data collection and manuscript writing.

Funding Statement

Children's Hospital of Philadelphia Foerderer Award-GRT-00001313. The study sponsors had no role in study design, data collection, analysis, or interpretation, writing of the manuscript, or decision to submit the manuscript for publication. SEV was supported by the National Cancer Institute under award numbers 1U01CA248226-01 and the DOD Peer Reviewed Cancer Research Program (W81XWH-21-1-0345, W81XWH 19-1-0668, W81XWH-21-1-0725), the Ohio Third Frontier Technology Validation Fund,

and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University and sponsored research funding from Pfizer. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Author Contributions

David A. Ostrowski and Joseph R. Logan were responsible for writing the first draft of the manuscript. Joseph R. Logan, Reiley Broms and Yong Fan were responsible for acquisition and curation of the dataset used for model training. David A. Ostrowski, Joseph R. Logan, Austin Thompson, Reiley Broms, Thomas Syphan, Maria Antony, Mandy Rickard, Lauren Erdman, Dennis Head, Jessica H. Hannick, Lynn L. Woo, Frederick D. Grant, Neeta D'Souza, Satish E. Viswanath, Chris A. Flask, Armando J. Lorenzo, Yong Fan, Gregory E. Tasian and John K. Weaver were responsible for the conception and design of the study, data analysis, interpretation of the data, and writing the final manuscript. Gregory E. Tasian and John K. Weaver provided clinical and overall oversight and supervision for the project. All

authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials

The datasets generated for this study are available on request to the corresponding author. A Github repository of the machine learning models utilized in generating the data presented in this manuscript, including patient summary data, is being developed and is published with this manuscript publication.

Ethics Approval

The studies involving human participants were reviewed and approved by CHOP Institutional Review Board (IRB#: 17-014098). Given the retrospective and de-identified nature of this study, the CHOP IRB approved a waiver of informed consent.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix A

TABLE A1. Permutation scoring of the relative importance of the input variables for the output predictions of the Clinical Model with 95% confidence intervals

Clinical feature	Relative weight
Differential function per unit volume of the affected kidney	0.104 ± 0.047
$t_{1/2}$ (minutes)	0.081 ± 0.041
Age at initial renal scan (years)	0.060 ± 0.026
Differential function per total volume of the affected kidney	0.009 ± 0.012

Note. The relative weight is a quantification of the importance of each input variable of the Clinical Model in determining the model's final prediction with greater relative weights corresponding to a greater influence in the model's output.

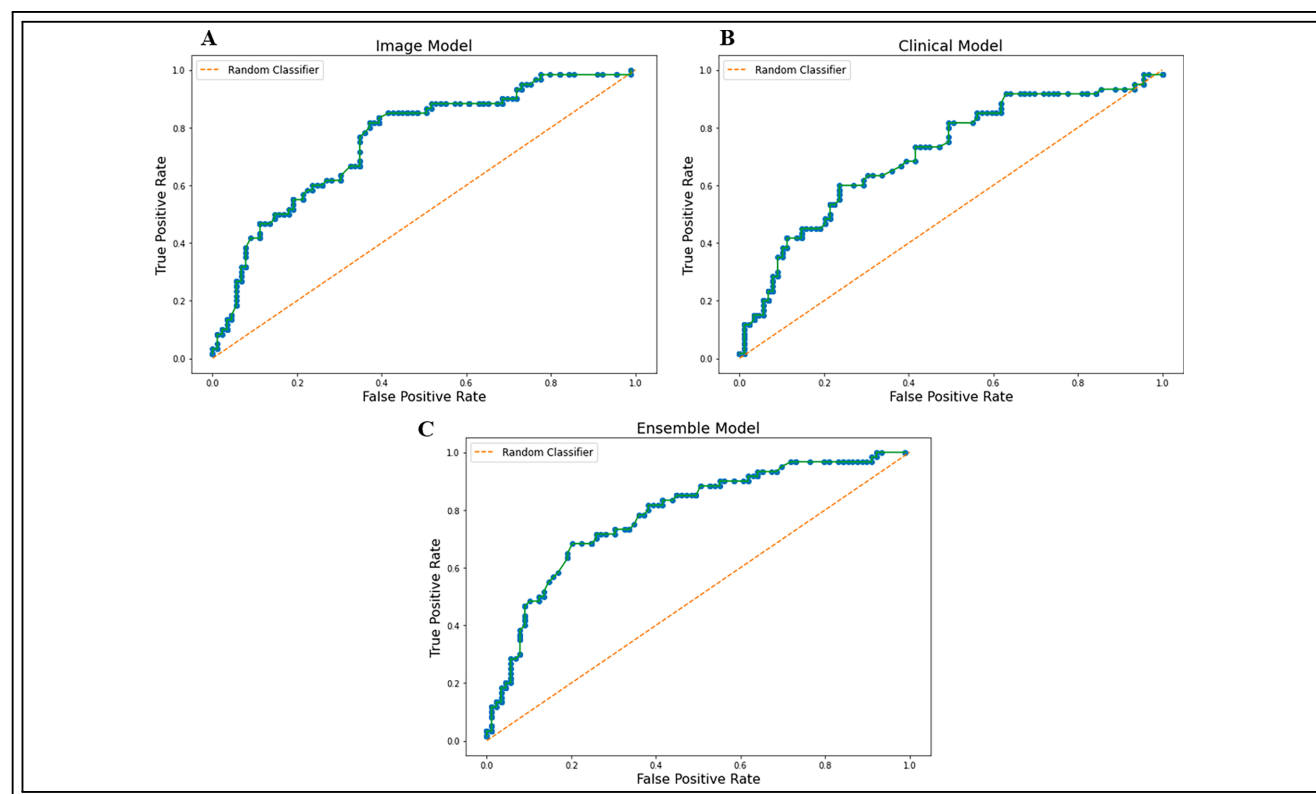


FIGURE A1. Receiver operating characteristic (ROC) curves of the three models at 6 years. (A) Imaging model with AUC 0.76 (95% CI: 0.72–0.81). (B) Clinical model with AUC 0.73 (95% CI: 0.70–0.76). (C) Ensemble model with AUC 0.80 (95% CI: 0.75–0.86).

References

- Woodward M, Frank D. Postnatal management of antenatal hydronephrosis. *BJU Int* 2002;89(2):149–156. doi:10.1046/j.1464-4096.2001.woodward.2578.x.
- González R, Schimke CM. The prenatal diagnosis of hydronephrosis, when and why to operate? *Arch Esp Urol* 1998;51:575–579.
- Braga LH, McGrath M, Farrokhyar F et al. Associations of initial society for fetal urology grades and urinary tract dilatation risk groups with clinical outcomes in patients with isolated prenatal hydronephrosis. *J Urol* 2017;197:831–837. doi:10.1016/j.juro.2016.08.099.
- Yang Y, Hou Y, Niu ZB et al. Long-term follow-up and management of prenatally detected, isolated hydronephrosis. *J Pediatr Surg* 2010;45(8):1701–1706. doi:10.1016/j.jpedsurg.2010.03.030.
- Hanna MK. Antenatal hydronephrosis and ureteropelvic junction obstruction: the case for early intervention. *Urology* 2000;55(5):612–615. doi:10.1016/s0090-4295(00)00460-x.
- Sussman RD, Blum ES, Sprague BM et al. Prediction of clinical outcomes in prenatal hydronephrosis: importance of gravity assisted drainage. *J Urol* 2017;197:838–844. doi:10.1016/j.juro.2016.09.111.
- Krill AJ, Varda BK, Freidberg NA et al. Predicting the likelihood of prolongation of half-time among infants with initially indeterminate drainage values: a single-institution retrospective study of 535 patients with ureteropelvic junction obstruction. *J Pediatr Urol* 2021;17(4):512.e1–512.e7. doi:10.1016/j.jpuro.2021.05.022.
- Kletter K, Nürnberger N. Diagnostic potential of diuresis renography: limitations by the severity of hydronephrosis and by impairment of renal function. *Nucl Med Commun* 1989;10(1):51–61. doi:10.1097/00006231-198901000-00008.
- 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 Council–The Society of Nuclear Medicine. *J Nucl Med* 1992;33(5):2047–2051. doi:10.1007/bf00860948.
- Karacalioglu O, Ilgan S, Arslan N et al. Unilateral temporary functional stasis in the upper urinary tract caused by a filled bladder on Tc-99m DTPA diuresis renography: a teaching case. *Ann Nucl Med* 2005;19(6):511–514. doi:10.1007/bf02985580.
- Majd M, Bar-Sever Z, Santos AI et al. The SNMMI and EANM procedural guidelines for diuresis renography in infants and children. *J Nucl Med* 2018;59(10):1636–1640. doi:10.2967/jnumed.118.215921.
- Wong J, Rickard M, Dos Santos J et al. Antenatal hydronephrosis: an updated review on postnatal care and management. *Curr Opin Pediatr* 2025;37(5):473–481. doi:10.1097/mop.0000000000001485.

13. Abdovic S, Cuk M, Cekada N et al. Predicting posterior urethral obstruction in boys with lower urinary tract symptoms using deep artificial neural network. *World J Urol* 2019;37(9):1973–1979. doi:10.1007/s00345-018-2588-9.
14. Scott Wang HH, Vasdev R, Nelson CP. Artificial intelligence in pediatric urology. *Urol Clin* 2024;51:91–103.
15. Yin S, Peng Q, Li H et al. Multi-instance deep learning of ultrasound imaging data for pattern classification of congenital abnormalities of the kidney and urinary tract in children. *Urology* 2020;142:183–189. doi:10.1016/j.urology.2020.05.019.
16. Kwong JCC, Khondker A, Kim JK et al. Posterior urethral valves outcomes prediction (PUVOP): a machine learning tool to predict clinically relevant outcomes in boys with posterior urethral valves. *Pediatr Nephrol* 2022;37(5):1067–1074. doi:10.1007/s00467-021-05321-3.
17. Weaver JK, Milford K, Rickard M et al. Deep learning imaging features derived from kidney ultrasounds predict chronic kidney disease progression in children with posterior urethral valves. *Pediatr Nephrol* 2023;38(3):839–846. doi:10.1007/s00467-022-05677-0.
18. Smail LC, Dhindsa K, Braga LH et al. Using deep learning algorithms to grade hydronephrosis severity: toward a clinical adjunct. *Front Pediatr* 2020;8:1. doi:10.3389/fped.2020.00001.
19. Lorenzo AJ, Rickard M, Braga LH et al. 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;123:204–209. doi:10.1016/j.urology.2018.05.041.
20. Cerrolaza JJ, Peters CA, Martin AD et al. Quantitative ultrasound for measuring obstructive severity in children with hydronephrosis. *J Urol* 2016;195:1093–1099. doi:10.1016/j.juro.2015.10.173.
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(3):847–852. doi:10.1016/j.juro.2017.09.147.
22. Weaver JK, Logan J, Broms R et al. Deep learning of renal scans in children with antenatal hydronephrosis. *J Pediatr Urol* 2023;19(5):514.e1–514.e7. doi:10.1016/j.jpuro.2022.12.017.
23. Smokvina A, Grbac-Ivanković S, Girotto N et al. The renal parenchyma evaluation: MAG3 vs. DMSA. *Coll Antropol* 2005;29:649–654.
24. Ritchie G, Wilkinson AG, Prescott RJ. Comparison of differential renal function using technetium-99m mercaptoacetyl triglycine (MAG3) and technetium-99m dimercaptosuccinic acid (DMSA) renography in a paediatric population. *Pediatr Radiol* 2008;38:857–862.
25. Nimmo MJ, Merrick MV, Allan PL. Measurement of relative renal function. A comparison of methods and assessment of reproducibility. *Br J Radiol* 1987;60(717):861–864. doi:10.1259/0007-1285-60-717-861.
26. Calle-Toro JS, Barrera CA, Khrichenko D et al. R2 relaxometry based MR imaging for estimation of liver iron content: a comparison between two methods. *Abdom Radiol* 2019;44(9):3058–3068. doi:10.1007/s00261-019-02074-4.
27. Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res* 2020;21:1–6.
28. Weaver JK, Martin-Olenski M, Logan J et al. Deep learning of videourodynamics to classify bladder dysfunction severity in patients with spina bifida. *J Urol* 2023;209(5):994–1003. doi:10.1097/ju.0000000000003267.
29. Lee JN, Kang JK, Jeong SY et al. Predictive value of cortical transit time on MAG3 for surgery in antenatally detected unilateral hydronephrosis caused by ureteropelvic junction stenosis. *J Pediatr Urol* 2018;14(1):55.e1–55.e6. doi:10.1016/j.jpuro.2017.08.009.
30. Piepsz A. Antenatal detection of pelviureteric junction stenosis: main controversies. *Semin Nucl Med* 2011;41(1):11–19. doi:10.1053/j.semnuclmed.2010.07.008.
31. Acker MR, Clark R, Anderson P. Gravity-assisted drainage imaging in the assessment of pediatric hydronephrosis. *Can Urol Assoc J* 2016;10(3–4):96–100. doi:10.5489/cuaj.3237.