A CLINICAL PREDICTION MODEL OF SURGERY AND OBSTRUCTION IN CHILDREN WITH ANTENATAL HYDRONEPHROSIS

DERIVATION AND VALIDATION OF CLINICAL PREDICTION MODEL OF SURGICAL RISK AND RENAL SCAN OBSTRUCTION IN CHILDREN WITH ISOLATED UNILATERAL ANTENATAL HYDRONEPHROSIS

By ROSEANNE FERREIRA DE FREITAS EUZEBIO, MD

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science in Health Research Methodology

MASTER OF SCIENCE (2023)	Health Research Methodology - Clinical Epidemiology McMaster University, Hamilton, Ontario
TITLE	Derivation and validation of clinical prediction model of surgical and obstruction risk in children with isolated unilateral antenatal hydronephrosis.
AUTHOR	ROSEANNE FERREIRA, MD
SUPERVISOR	LUIS BRAGA, MD, MSc, PhD
NUMBER OF PAGES	vxi, 116

LAY ABSTRACT

Hydronephrosis (HN) is a kidney condition prevalent in 1-5% of pregnancies, typically detected through third-trimester ultrasounds. Although 70% of cases resolve spontaneously, the remaining 30% are at higher risk of renal failure if unaddressed. Traditional scoring systems have depended on renal scans, which expose patients to radiation and suffer from inconsistency across healthcare facilities. Using data from pediatric patients with HN, this study identified key predictive factors for surgical management such as severe dilation of kidneys associated with thinning, affected kidney length more than 12% longer than healthy kidney and anteroposterior diameter of the renal pelvis > 13mm. A new scoring system based solely on renal ultrasound parameters was formulated. This system establishes different risk categories, which have been validated for use in daily clinical practice. This offers urologists a reliable, radiation-free tool to gauge individual surgical risks and provides a foundation for tailored patient counselling.

ABSTRACT

Introduction: There is no single agreed-upon tool to forecast hydronephrosis obstruction and surgical risk in the literature. Previous scoring systems have relied on renal scan parameters for this task, which are not performed or read consistently across hospitals. We aim to determine the utility of a hydronephrosis obstruction early scoring system in identifying severe UPJO-like hydronephrosis suggestive of surgery.

Methods: We reviewed a prenatal hydronephrosis database (2008-23) to identify patients with unilateral UPJO-like hydronephrosis. Patients with vesicoureteral reflux (VUR), megaureter, or other anomalies were excluded. We collected data on the first ultrasound and surgical outcome for the Society for Fetal Urology (SFU) grade, anteroposterior pelvic diameter (APD), and percent difference in renal length – the renal length index (RLI). I undertook a multivariable logistic regression to determine the predicted value of these sonographic measurements. Model development and internal validation were performed, and model performance (calibration and discrimination) was reported.

Results: Of the 465 UPJO-like hydronephrosis patients, 76.1% (354) were male and had a mean age at baseline of 3.8 ± 3.8 months. 34.8% (162) had an obstructive renal scan and 27.7% (129) underwent pyeloplasty. The developed model to predict surgery using the aforementioned variables with excellent model performance (AUC 0.92). A simplified model using SFU 3-4, APD > 13mm and RLI >12% also predicted surgery with great model performance (AUC 0.89) and improved clinical user-friendliness.

Conclusion: Our clinical predictive model can accurately predict surgery in both derivation and validation cohorts of patients with antenatal UPJO-like hydronephrosis. This non-invasive tool can greatly aid in patient counselling.

ACKNOWLEDGEMENTS

First and foremost, I would like to dedicate this thesis to my grandmother Ruth (*in memoriam*). Though you are no longer with us, your enduring guidance, relentless drive, and resilient strength have left an indelible mark on my life. Your wisdom and love continue to inspire me each day and have been a profound influence on my academic journey. I aspire to uphold the values you instilled in me in all my future endeavours. To my husband and mother, I am forever grateful for the unwavering support, encouragement and laughs that uplifted me during this journey.

I would also like to extend my deepest gratitude to my supervisor, Dr. Braga. Your insightful feedback and mentorship have been invaluable to me throughout the course of this research. Your expertise and guidance have not only contributed to the quality of this work but have also facilitated my growth as a researcher and physician.

To my dear Dr. Farrokhyar, your constructive feedback, encouraging words, and consistent support have been a pillar of strength in times of uncertainty. Your insights have been indispensable in shaping both this thesis and my perspective on the true beauty of statistics.

Lastly, I would like to thank my defence committee, friends, family, and colleagues who have contributed in uncountable ways to the completion of this work. Your encouragement, both spoken and unspoken, has made this journey worthwhile.

TABLE OF CONTENTS

LIST OF TABLES/FIGURES

Tables

Figures

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AIC	Akaike's Information Criterion
ANOVA	Analysis of Variance
APD	Anteroposterior pelvic diameter
AUC	Area under the curve
CUA	Canadian Urological Association
CITL	Calibration in the large
CI	Confidence interval
df	Degrees of freedom
DFR	Differential renal function
DMSA	Dimercapto succinic acid
EAU	European Association of Urology
GFR	Glomerular filtration rate
HERO	Hydronephrosis early risk of obstruction
HI	Hydronephrosis Index
HiREB	Hamilton Integrated Research Ethics Board
HL	Hosmer-Lemeshow
HN	Hydronephrosis
HSS	Hydronephrosis Severity Score
IQR	Interquartile range

LR	Likelihood ratio
MAG3	Mercaptoacetyltriglycine
MAG3-SOS	MAG-3 Suspected obstruction scoring system
MAR	Missing at Random
PHAR	Parenchyma to hydronephrosis area
PPS	Pyeloplasty Prediction Score
PVP	Positive Predictive Value
RLI	Renal length index
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SE	Standard Errors
SFU	Society for Fetal Urology
T ½ time	Time to 50% clearance of tracer
UPJO	Ureteropelvic junction obstruction
USD	United States dollar
VCUG	Voiding cystourethrogram
VIF	Variance Inflation Factor
VUR	Vesicoureteral reflux

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Roseanne Ferreira, declare I am the primary author of all chapters included in this thesis. I am also primarily responsible for the statistical analyses, interpretation, and reporting of data under the supervision of Dr. F. Farrokhyar, Dr. L. H. Braga and Dr. R. Chanchlani. Patient data were collected by a qualified team led by Dr. L. H. Braga, including myself. To the best of my knowledge, this thesis does not infringe upon any existing copyrights.

CHAPTER 1: INTRODUCTION

Hydronephrosis background and burden

Hydronephrosis (HN) is a condition characterized by dilation of the renal pelvis and calyces where one or both kidneys swell due to improper drainage of urine. It is one of the most common urinary system abnormalities diagnosed on ultrasound during the third trimester[1], affecting 1-5% of pregnancies[2]. Ureteropelvic junction obstruction (UPJO) is one of the most common etiologies of HN in children with an incidence of one in 500 live births screened on prenatal ultrasound[3]. Despite having a high-resolution rate, where up to 70% of the patients resolve spontaneously[4–8], hydronephrotic patients are four times more likely to develop end-stage renal disease[9] and are at risk of requiring renal replacement in early adulthood[10].

Historically, diagnosis of UPJO and hydronephrosis was primarily a postnatal event, often occurring after the onset of clinical symptoms or when renal function had already been compromised. The advent of antenatal ultrasonography has radically altered this landscape and patients are being diagnosed at earlier age[11]. Prenatal detection of the condition has not only led to proactive management strategies but also offers the potential to intervene before any irreversible renal damage has occurred[12]. However, this early diagnostic capability is not without its challenges. While it enables more immediate action, discerning which cases truly require surgical intervention versus those that may resolve spontaneously remains a complex task[13].

The enduring clinical conundrum in managing antenatal UPJO hydronephrosis lies in identifying those cases of persistent obstruction that will indeed necessitate surgical intervention. Distinguishing these from instances of transient, non-obstructive hydronephrosis is not straightforward and is subject to both diagnostic and interpretative uncertainty[11,13]. Several diagnostic modalities, including renal scan and serial ultrasonography, have been proposed to aid this differentiation[14]. However, no single method has emerged as a definitive gold standard. Consequently, clinicians often find themselves navigating a complex decision-making process clouded with diagnostic ambiguities and institutional protocol discrepancies.

This diagnostic uncertainty has significant implications. On one hand, failure to promptly identify and surgically correct genuine obstructions can lead to irreversible renal damage. On the other, overly aggressive surgical management may expose patients to unnecessary risks and healthcare systems to undue financial burdens. Moreover, the absence of a consensus on the best agreed-upon method to define renal obstruction and treatment approaches can result in varied clinical practices, further complicating the landscape of UPJO management[11].

Moreover, antenatal diagnosis has dramatically impacted healthcare costs and resource allocation[15–17]. While early diagnosis through serial ultrasonographic monitoring, specialized pediatric consultations, and potential surgical interventions offer the benefit of preserving renal function and improving long-term outcomes, these procedures come with significant financial ramifications[17]. This burden is felt both at the systemic level and individual level, where moderate to severe patients had 1.5 and 2 times significantly higher

number of imaging in the first year of life, respectively[16]. During a cost-analysis for resource utilization of patients with non-refluxing HN, Akhavan and colleagues identified that the resource utilization for patients less likely to undergo surgery was less cost-effective[17]. The overall cost (2014 USD) of identifying a single surgical case was infinite for grade SFU 1, $37,599.80 for SFU 2, $11,741.19 for SFU 3 and $2,142.02 for SFU 4[17], highlighting the inadequate allocation of funds associated with the uncertainty around the surgical management of patients.

However, the burden is not purely financial; the emotional toll on families is considerable and often overlooked[18]. A prenatal diagnosis can plunge parents into a state of anxiety, exacerbated by the uncertainty of postnatal prognosis. This compounds the strain on families, who identified the logistic aspect of the decision-making process to be one of the major barriers to guiding treatment decisions for their child with UPJO. They highlighted the importance of having accurate information targeting patients' individualized risk to better inform their decision[18]. Addressing both these financial and emotional aspects is also vital for a comprehensive, patient-centred care model.

Prognostic factors of hydronephrosis severity

Previous literature has explored abundantly available radiological measures of morphology and renal function to identify factors associated with HN progression and surgical intervention[11,13,14]. Understanding the nuances of renal anatomy and physiology is pivotal when considering prognostic factors for evaluating the severity of the disease. With distinct sub-organs such as the renal pelvis, calyces, medulla, and cortex, each of which responds

differently to HN, their respective anatomical and functional metrics offer a spectrum of indicators[13]. For instance, the compliance of the renal pelvis, particularly in infants, can serve as a protective anatomical response, diminishing the risk of renal parenchymal damage[13]. On the other hand, dilation of the calyces can imply a higher severity of disease compared to renal pelvic dilation alone, due to its lower compliance[19]. These are functional adaptations to protect parenchyma from damage.

The most crucial is the assessment of renal parenchyma quality, notably its thickness and appearance[13]. In this regard, long-lasting cortical thinning, or reductions in renal size correlate with irreversible renal disease and functional loss, which underscores disease severity[14]. Specific imaging findings like hyperechoic parenchyma and loss of corticomedullary differentiation further indicate severe renal damage, correlating well with tubular atrophy and interstitial inflammation[20].

To enable clinicians to identify candidates for surgical intervention more accurately, the selection of prognostic factors should include imaging metrics that translate these anatomical and physiological individualities, that are objective and ideally subject to minimal intra- or inter-observer variation.

The Anteroposterior pelvic diameter

The Anteroposterior renal Pelvic Diameter (APD) is an objective ultrasound-based parameter. Measured in the mid-renal transverse plane, APD quantifies the width of the renal pelvis, a quantitative estimate for assessing the degree of dilation and informing

treatment planning across both prenatal and postnatal life stages[21]. In that manner, fetal APD has been the most extensively studied predictor of postnatal renal obstruction[14].

APD has been suggested as an independent predictor of pyeloplasty, and multiple studies have delineated its efficacy in predicting surgical intervention. For instance, Dias et al. found that prenatal and postnatal APD cut-offs of 16 mm and 18 mm, respectively, were highly predictive of surgery[22]. Sharifian et al. supported a 15 mm cut-off with AUC 0.92 for surgery[23], while Burgu et al. suggested that a diameter of less than 20 mm could indicate preserved differential renal function[24]. These cut-off points from real-world data offer actionable insights for clinicians, particularly in estimating the necessity for surgical intervention.

However, it's crucial to recognize that APD's reliability is contingent upon a variety of factors. These include the patient's hydration status, bladder volume, and positioning during imaging[11,25]. Additionally, anatomical considerations, such as whether the renal pelvis is extrarenal or intrarenal, can affect the predictive accuracy of APD measurements. These variables have the potential to alter the perceived extent of parenchymal damage or to overestimate the severity of obstruction[13]. In light of these considerations, understanding APD's limitations is essential for maximizing its clinical utility. Variabilities in APD measurement techniques call for a systematic approach to patient preparation and imaging techniques when measuring and comparing sequential ultrasounds from the same patient.

Driven by the lack of consensus in accurate location for measurement or information on the best practices for APD measures, a recent study by Hodhod and colleagues outlined six technical variations in APD measurement[25]. All variations displayed strong correlations

with pyeloplasty outcomes, with discrimination indices ranging from 0.89 to 0.91[25]. They found supine versus prone position measurement to have minor differences, and despite intra-renal APD being smaller than its extrarenal counterparties, they were still associated with pyeloplasty. Their results reiterate the robustness of APD to forecast hydronephrosis outcomes. Furthermore, the development of specialized criteria using APD such as the PI-APD by Braga et al[26] highlights the evolving utility of this diagnostic tool, indicating its importance not only for immediate treatment decisions but also for dynamic monitoring and long-term evaluation of post-surgical improvement.

Despite these important findings, APD alone cannot fully assess other significant facets such as abnormal renal morphology or parenchymal integrity, highlighting the need for a more comprehensive evaluation strategy that may include additional parameters such as degree of calyceal and pelvic dilation.

The Society for Fetal Urology grading system

The Society for Fetal Urology (SFU) grading system is a widely used instrument in current urological practice, assisting healthcare providers in forming an initial diagnosis and offering ongoing continuity of care. Since its introduction in 1993, SFU has become a cornerstone in the evaluation of hydronephrosis, particularly in pediatric cases and used by pediatric urologists[27]. The system uses a semi-quantitative scale, ranging from grades 0 to 4, that relies on renal morphology such as renal pelvis dilation, calyceal changes, and parenchymal thinning[8,27]. This structured severity stratification aids both in risk assessment, clinical decision-making and use in clinical trial criteria[28,29].

The predictive power of SFU is substantiated by real-world data studies, particularly in determining the need for surgical intervention. A cohort study by Braga et al. involving 501 UPJO-like patients demonstrated that the rate of pyeloplasty in patients with SFU grades 3 and 4 was significantly higher than in those with SFU grades 1 and 2, specifically 32% compared to 2%[30]. Additionally, higher SFU grades were found as independent predictors of deteriorating renal function, often requiring timely surgical action to prevent further renal damage[7,8,27,30]. There is an association between preserved differential renal function and SFU grade 3 hydronephrosis, while SFU grade 4 was linked to declining renal function and hence, increased surgical intervention[11]. These correlations are pivotal in pediatric cases, where they shape both initial and ongoing management strategies.

However, the SFU is not without its limitations. Primarily anatomical in focus, the SFU grading overlooks functional aspects, such as differential renal function, possibly leading to an oversimplification of complex cases[31]. Additionally, attempts to augment its precision have led to alternative grading proposals. For example, Sibai et al. have suggested a sub-classification within SFU grade 4 to differentiate between segmental and diffuse cortical thinning[31]. This nuanced approach could potentially improve this grading system's predictive accuracy, though further validation is warranted.

While the SFU grading system serves as an indispensable tool in modern urological practice, its limitations do not support its use as a single tool for severity evaluation. Its role in both initial diagnostic workup and ongoing treatment underscores its importance but also highlights the necessity for integrating functional parameters to enhance predictive accuracy and individualized patient care.

Renal length

The pursuit of reliable prognostic markers in the diagnosis and management of HN has been subject to ongoing clinical and basic research. Notably, renal length emerges as a parameter worth exploring, given its promising role in understanding the severity of hydronephrosis and renal function deterioration[32,33]. It is a quantitative measurement of the mid-polar longitudinal size of the kidney.

Traditionally, the diagnosis and therapeutic interventions for hydronephrosis have relied on many parameters, yet the dynamic relationship of severity between kidneys has been mostly explored by nuclear medicine and its sonographic assessment remains subject to debate. The work of Luke et al. adds a valuable dimension to this conversation by focusing on relative renal length, quantifying how much smaller the healthy kidney is in relation to the affected one [34]. Their study shows that the unaffected kidney grew more in both pyeloplasty and non-pyeloplasty groups (11.4% vs 2.1%, $p<0.001$) and patients undergoing pyeloplasty had significantly higher discrepancy between the kidney sizes, thereby suggesting a significant difference in the sizes of the kidneys, especially among those undergoing surgical intervention; namely, 6.3 vs 5.5cm, $p<0.001$ for pyeloplasty groups and 5.8 vs 5.6cm, $p<0.001$ for the non-pyeloplasty ones [34]. This underscores the idea that renal length could be a pivotal indicator for surgical intervention.

The advantages of employing renal length as a diagnostic marker extend further. Utilizing the ratio between renal lengths produces a unitless measure, where each individual serves as their control. This seamlessly circumvents the discrepancies that arise when comparing

renal size across different ages or height[35,36], making it a particularly useful tool for longitudinal assessment in pediatric populations.

Animal models have also reinforced the relevance of renal length as a parameter. Studies on rabbits indicated that renal size is particularly instructive in evaluating the true function of obstructed kidneys, as well as their potential for recoverability[32]. A marked decline in baseline Glomerular Filtration Rate (GFR) and a significant expansion in renal size were predictors of advanced obstruction and limited renal reserve compared to a smaller increase in renal size during early obstruction phases[32].

Additionally, our understanding of renal length in the context of healthy children adds another layer to its potential utility. It is generally observed that the left kidney tends to be larger than the right one [37–39]. Khazaei and colleagues went further to identify that discrepancies in renal lengths could effectively predict abnormal dimercaptosuccinic acid (DMSA) scans, the gold standard for parenchymal damage [33]. For children under 4 years old with a longer right kidney, discrepancies greater than 6mm between renal lengths predicted abnormal DMSA scan (differential renal function (DRF) <45% and/or a scar) with 36% sensitivity, 100% specificity and 100% positive predictive value[33]. Specific renal length discrepancy cut-offs were associated with high specificity and positive predictive value, rendering them potentially useful in clinical practice [33]. Lastly, a study from Li et al. identified that a score including the absolute value of renal length difference associated with other sonographic measurements identified higher discrepancy values to be associated with a higher likelihood of surgery[40].

The relationship between renal length and renal obstruction requiring surgery is still largely unexplored and yet shows promising results. Its inherent advantages of being unitless, allowing for intra-individual comparison, and its predictive utility in both animal models and pediatric populations, position it as a noteworthy metric for further exploration in clinical settings.

The Need for a Comprehensive Prognostic Mode

The clinical management of hydronephrosis, particularly in determining the need for surgical intervention, has been a subject of ongoing debate and inquiry within the urological community. Despite a wide range of diagnostic tools at our disposal, including all measurements recommended by the European Association of Urology (EAU) to be evaluated during follow-up—such as APD, calyceal dilation, kidney size, function residual urine—there remains a significant gap in the establishment of a universally accepted, reliable prognostic metric[29]

The limitations of current methodologies are evident and applicable to other imaging modalities. Renal scan findings, DRF and T 1/2 time of renal drainage, both valuable in HN evaluation, suffer from inconsistencies and interpretational challenges. The DRF drop as a surgical indicator varies between studies, with thresholds ranging from 5% to 10%[7,41,42]. Additionally, some authors recommend against using T1/2 time to indicate the need for surgery since renal function may be overrepresented in the child's large hydronephrotic kidney [5,41,43,44]Adding to this complexity is the lack of standardization in renal scans, which differ across institutions and often lead to variable interpretations[43].

Given these shortcomings, there's a compelling need for a comprehensive, standardized, and non-invasive prognostic model. This is especially relevant considering that approximately 85% of surgical decisions for UPJO patients already rely on ultrasonographic parameters[13]. Therefore, this thesis aims to develop a prognostic score focusing solely on ultrasonographic parameters. By integrating SFU, APD, and renal length index, this project offers a robust, multi-dimensional approach that leverages physiologically relevant parameters predictive of renal function obstruction and deterioration. The hypothesis driving this work is that a predictive model based on these three measures on baseline ultrasound can accurately forecast the need for surgical intervention in cohorts with unilateral UPJO hydronephrosis.

Thesis objectives

This study aims to answer the following objectives:

1) To describe the renal length index (RLI) and evaluate its ability as a prognostic factor for surgery.
2) To derive and internally validate a clinical prediction model of surgery in patients with unilateral isolated antenatal hydronephrosis using the first post-natal renal ultrasound measurements of SFU, RLI, APD and potential patient characteristics.
3) To derive and validate a simplified scoring system using the first post-natal ultrasound measurements of SFU, RLI and APD to facilitate usability in daily clinical practice.

CHAPTER 2: METHODS

Study design

This is a retrospective cohort study for prognostic model development. Data was analyzed from a prospectively collected cohort database from a single high-volume pediatric urologist from a tertiary academic pediatric hospital. This study was reviewed and approved by the Hamilton Integrated Research Ethics Board (HiREB approval number 14448).

The prenatal hydronephrosis database is comprised of de-identified patient-level data from children with antenatal diagnosis of hydronephrosis and their outcomes. Information on these children is updated at every clinical follow-up appointment. Frequency of clinical visits, thus data collection points, varied according to the severity of hydronephrosis. Data analyzed comprised of patients receiving urological care between January 2008 to February 2023.

Eligibility criteria

Children with the antenatal diagnosis of hydronephrosis who had at least 6 months of follow-up were included in the study. We excluded patients 1) whose hydronephrosis etiologies were vesicoureteral reflux (VUR), ureteral atresia, or hydroureteronephrosis; 2) Bilateral cases; 3) with anomalies on the affected or contralateral kidney (ureterocele, ectopic kidney, horseshoe kidney, posterior urethral valves, multicystic dysplastic kidney disease, simple duplication, ectopic ureter, solitary kidney, neurogenic bladder, spina bifida

or prune belly); 4) Patients undergoing other surgery than pyeloplasty to manage the hydronephrosis (Figure 1).

Data collection and Monitoring

Postnatal ultrasound

Data from variables of interest were retrieved from the first (baseline) renal ultrasound performed in our institution. Renal ultrasounds were performed using institutional protocol as follows: 1) no pretest hydration, 2) patient positioned at supine position, 3) Pre-void and post-void measurements to confirm an empty bladder, and 4) as soon as feasible after 48 hours after birth. All ultrasounds were performed by technicians specialized in pediatric renal bladder ultrasound, who were assigned to the pediatric Urology service at McMaster Children's Hospital. Measurements retrieved from these ultrasounds were evaluated by fellowship-trained radiologists who have experience in pediatric renal bladder ultrasound.

Measurement of sonographic variables

Transverse APD was calculated by measuring the distance between parenchymal borders in the mid-section of the renal hilum. Extra and intra-renal measurements of APD were taken and the larger one was recorded[45]. For the renal length, the measurement between the most distant point on the upper and lower pole was taken for contralateral and ipsilateral kidneys in the longitudinal view of the US. The SFU grade was measured following the SFU grading system based on dilation of calyces, renal parenchyma, and renal pelvis[27].

Patient follow-up

Patient follow-up was done according to their hydronephrosis severity following institutional and Canadian Urological Association (CUA) guidelines[2]. Patients underwent a voiding cystourethrogram (VCUG) to evaluate for VUR. Patients with high-grade hydronephrosis (SFU grades 3 or 4, APD > 15mm) were followed at the clinic every 3 months for the first year, every 6 months until the second year and yearly after that. Patients with low-grade hydronephrosis (SFU 1 or 2, APD < 10mm) were followed every 6 months until the second year and yearly after that. Research personnel updated the database during each follow-up appointment. Data on patients meeting eligibility criteria were audited by myself to ensure data accuracy.

Statistical Analysis Plan

Descriptive statistics

Baseline characteristics were described using descriptive statistics. The normality of distribution was assessed visually by plotting histograms and confirmed by the Shapiro-Wilk test. Continuous data was described as mean ± Standard deviation (SD) or median and interquartile range (IQR) given their normality of distribution. T-tests or Wilcoxon rank sum were used to identify differences between the cohorts, and for comparison between continue variables with three or more categories; ANOVA or Kruskal Wallis tests were used, followed by a post hoc analysis with Bonferroni test. Categorical data was described as counts and percentages. Fisher's exact test or chi-square (when Fisher was not computationally supported by the software) was used to describe differences in categorical

variables between derivation and internal validation cohorts. A two-sided p-value <0.05 was defined as statistically significant.

Data splitting

Data splitting methods are one common approach to developing and validating prognostic models[46]. The dataset was divided into 2:1. Two-thirds were a derivation cohort, used for model development, and the final third (validation cohort) used for model internal validation. Data was divided into thirds by using stratified random data splitting to ensure that each dataset was representative of the different HN severity groups and outcome of interest. The Strata of choice were baseline SFU and surgical status. This method was selected since the different HN severity distributions are generally imbalanced in the population, with mild HN being more prevalent. Also, it was a strategy to improve model performance as it would be derived and validated in a sample representative of the real-world prevalence found between severity groups, such as it is in daily clinical practice where the model is being targeted to[47]. Moreover, it would help reduce overfitting and improve the generalization ability of the model[46]. We have considered k-fold cross-validation as an option since it would allow us to train the model in different folds of the dataset, thus outputting a more accurate sample error estimation of the model performance. However, this method would not be sufficient to control overfitting in our small sample size with a low outcome rate, thus would have produced biased performance estimates[48].

Outcome variable

The outcome variable was surgery via open or laparoscopic pyeloplasty. Surgery (pyeloplasty) was recommended for patients with HN worsening defined[2] by 1) an increase in APD with or without change in SFU grade on repeated ultrasounds, or 2) deterioration of differential renal function > 10% on repeated diuretic renography, or 3) renal function < 40% associated with an obstructive curve (ascending or flat), or 4) worsening of hydronephrosis with T ½ > 30 min, or 5) development of symptoms (febrile urinary tract infections, stones, sepsis).

Development of clinical prediction model of surgery

I performed a multivariable logistic regression analysis to develop a prognostic model where surgery was the dependent variable (Table *1*). The selection of independent variables for the model was informed by a review of the literature and relevant clinical reasoning. These variables were included in their original units.

The model was derived using a backward elimination strategy. This method involves starting with all candidate variables and removing the least significant one at each step until all variables remaining in the model are statistically significant at a specified level. In this case, the cut-off value was set at $p>0.20$ for removal from the model. Nevertheless, based on clinical rationale and the comprehensive evaluation of hydronephrosis parameters, the minimal model was stipulated to include APD, SFU grade, and RLI.

I tested continuous predictors for the linearity assumption using the Box-Tidwell transformation and visually inspecting the scatter plot between predictor and logit values.

In the case of a non-linear relationship, a restricted cubic spline with 4 knots was planned for the variable. Log transformations were not considered due to the high value placed on the model's interpretability and usability for clinicians. This approach enabled me to exclude non-significant variables, thereby simplifying the model and enhancing its interpretability for clinical practitioners.

Independent variables included in the model

APD, SFU, RLI, laterality, sex and age, were used as independent variables when deriving the logistic model, and a minimum model, including APD, SFU, and RLI were used.

Sample size and event rate

The total event rate for the dataset meeting eligibility criteria was 129 (27.7%). The data splitting method selected retrieved a total of 86 (27.4%) event rate for the derivation cohort and I evaluated a total of 6 covariates, 3 included in the final model. This resulted in 29 events per covariate in the final model. This meets the general rule of thumb from simulation models of at least 10 to 15 events per predictor variable to maintain model stability and avoid overfitting[49]. Additionally, interactions and multicollinearity between predictors were evaluated to prevent inflated standard errors and unstable parameters thus masking the true relationship between predictors and outcomes. If variables were highly correlated (variance inflation factor (VIF) > 5) the least significant one was removed from the model[50].

Calibration and discrimination

Calibration and discrimination are fundamental measures to assess the performance of reliability, clinical utility, and applicability of the model's predictions[51]. Calibration compares the agreement between the predicted probabilities of the present model to the observed outcomes. It was initially assessed by a Hosmer-Lemeshow test, followed by evaluation of a calibration plot - a graph plotting the proportion of observed patients undergoing surgery against the model's predicted probability. This reflected the model's accuracy across the range of probabilities. A slope of 1 would mean complete agreement between predicted and observed proportions, in addition to a calibration in the large (CITL) – the intercept of this plot – of 0 would suggest accurate average estimates. For a robust assessment, a locally weighted scatterplot smoothing (Lowess) line facilitated evaluation of the alignment across the range of observed and predicted probabilities.

Discrimination evaluated the model's ability to differentiate between patients who had the outcome (surgery) and those who did not. Concordance statistics (C-statistics) and area under the receiver operating characteristic (ROC) curve were used. C-statistics values closer to 1 indicated better discrimination performance.

Model validation

After the model was developed with the derivation cohort, it underwent internal validation with the remaining third of the dataset. This step was essential to demonstrate the derived model presented reliable estimates within the same population they were derived. A bootstrapping resampling technique with 1000 samples was performed as a sensitivity

analysis to evaluate how accurate the reported C-statistics for model discrimination performance were for the same population. New calibration curves and discrimination using bias-corrected C-statistics were reported for the validated model[51].

Hydronephrosis Early Risk of Obstruction (HERO) model: development of clinical prediction model using categorical variables

A model was created categorizing the continuous variables or using the original ordinal values of independent variables. The goal was to increase the user-friendliness of this prediction model and to facilitate implementation in clinicians' daily practice. Each variable presented 4 categories to align with SFU categories, minimize the amount of information loss from categorization, and align with previous work from our team[40].

Categories for ordinal variables, such as SFU, were kept on their original values. Continuous variables such as APD and RLI were dichotomized based on ROC curve analysis and clinically relevant values. Cut-off points were selected based on the sensitivity and specificity of plotted values, where a value with higher sensitivities and specificities (Youden's J Index) was preferred. Cut-off values with decimal points were rounded up to the next integer and clinically significant value. These cut-off values were later described regarding the variable distribution.

The simplified model calibration and discrimination were reported to compare with the model using variables on their original structure. Categorized variables were attributed scores given logistic regression diagnostics to capture the effect and predictive power of each category. The simplified HERO model values were attributed as a ratio to their log-odds and a nomogram was fitted.

Dichotomous model using categorical variables

Another simplified model was created by dichotomizing the independent variables as an initial way of identifying patients at higher risk of surgery. The goal was to increase the user-friendliness of this prediction model and to facilitate implementation in clinicians' daily practice as a triage and referral tool for non-specialists.

All independent variables such as SFU, APD and RLI were dichotomized based on ROC curve analysis and clinically relevant values. Cut-off points were selected based on the sensitivity and specificity of plotted values, where a value maximized sensitivities and specificities – their Youden's J Index. Cut-off values with decimal points were rounded up to the next integer.

The simplified dichotomized model calibration and discrimination were reported following the same methodology as the HERO score. Dichotomized variables were attributed scores given logistic regression diagnostics, namely as a ratio of their log-odds, to capture the effect and predictive power of each category.

Sensitivity analysis

Weighted versus unweighted HERO score

Developing a simplified score is essential to increase usability by clinicians during everyday practice by allowing seamless calculation of individual prognosis during patient counselling. For the simplified HERO score, I have opted to assign values using weights derived from the logistic regression log-odds and nomogram to capture the individualities from the variables in predicting outcomes. However, I anticipated that having different values for each variable category could add complexity to the score, thus, reducing usability. Therefore, I decided to compare model calibration and discrimination with an unweighted HERO score where equal weights were assumed and each predicting variable ranged from 0 to 4, with 1 point scored for each level of increasing severity to a total of 12 points. Discrimination for this sensitivity analysis was evaluated by assessing C-statistics and calibration by the Hosmer-Lemeshow test.

Surgery as a surrogate of obstruction

The best fitting simplified score values were utilized to evaluate its ability to predict renal obstruction on the renal scan. Patients with severe HN identified on US underwent a Technetium-99m (99mTc) mercaptoacetyltriglycine (MAG3) diuretic renography - a renal scan - to identify functional significances of urine transport. The exam was performed under standardized circumstances following guidelines on pre-procedure hydration and

catheterization[52] from McMaster Children's Hospital protocol. Renal scan parameters, namely the shape of the contrast drainage curve and half-time contrast excretion (t ½) are the most used surrogate outcomes for renal obstruction in clinical practice. Renal scan obstruction was defined as the presence of t½ ≥ 30 minutes or a post-diuretic obstructive drainage curve (Ascending, flat or dynamic obstruction)[53].

Thus, a secondary model was undertaken using a multivariable logistic regression using the same independent variables included in the best-fitting simplified model with the presence of obstruction on a diuretic renography as the dependent variable. Calibration was performed by the Hosmer-Lemeshow goodness of fit test followed by confirmation on a graphic comparison between observed and expected probabilities of outcomes. Discrimination performance derived from C-statistics and area under the ROC curve.

This step evaluated the predictive value of best-fitting simplified model to renal scan obstruction.

Missing data

The database used has been frequently audited and evaluated for completeness and accuracy of data by PI's research personnel. Should missing data be <5% for all variables, the dataset will be used in its entirety with no corrections. However, if missing data was greater than 5%, the reason for missingness would be considered as missing at random (MAR) and multiple imputations would have been utilized.

Statistical analyses

Stata 18BE (College Station, Texas) was used for statistical analysis. Significance was defined as $p<0.05$ and all tests were two-sided.

CHAPTER 3: RESULTS

Baseline characteristics

The Prenatal hydronephrosis database comprised of 1,136 patients. 407 were excluded due to etiologies other than UPJO-like, 162 for bilateral disease, 64 due to the presence of anomalies, and 38 due to surgical management other than pyeloplasty (Figure 1). The remaining 465 patients with isolated unilateral UPJO were considered for analysis.

The average age at baseline ultrasound was 2.8 (IQR 1.4–4.8) months, with males comprising 76.1% (354) of the cohort. HN was found predominantly on the left side (76.3%, 355). The severity of HN was distributed evenly in the cohort – 239 (51.4%) patients were SFU 1 or 2 while 226 (48.6%) were SFU 3 or 4. Hydronephrotic kidneys had an average APD of 11.2 ±7.8mm and a median length of 6.0mm (IQR 5.5-6.6mm)(Table 2).

Most patients (303, 65.2%) had a non-obstructed renal scan, with a drainage half time (t ½) less than 30 minutes in 82.8% (385) of patients and a descending drainage curve in 65.8% (306). Pyeloplasty was performed in 27.7% (129) of patients, while the rest (72.3%, 336) experienced spontaneous resolution during follow-up (Table *2* - Patient characteristics). Patients who underwent surgery were primarily SFU 3 and 4 (36.4%, 47 and 58.9%, 76 respectively), and had a significantly higher (APD 19.2±9.2mm vs. 8.1±4.2mm, $p<0.001$) (Table 3).

The Renal Length Index (RLI)

The renal length index was distributed normally on visual inspection (Figure 2) and had a median value of 7.0% (IQR 1.5-14.9%). RLI varied significantly across SFU categories ($F(3, 461) = 85.53$, $p < 0.0001$). Notably, the mean RLI values for SFU 1, 2, 3 and 4 categories were respectively 2.0±10.1%, 2.1±9.9%, 9.9±8.6%, 21.9±11.2%. Lower RLI mean values were observed in the less severe hydronephrosis category, SFU 1, whereas the most severe category, SFU 4, had the highest mean RLI. Regarding pairwise comparisons, except for SFU 1 and 2, all other categories demonstrated statistically significant differences in RLI, validated by Bonferroni-adjusted p-values less than 0.001. RLI means were not significantly different between HN laterality (left, 7.7±12.1% vs right, 9.0±12.7%, $p=0.328$).

In the context of renal scans, a stark contrast was found between obstructed and non-obstructed patients. Those without an obstruction had a mean RLI of 3.7±10.0%, while the mean value for the 162 obstructed patients was four times higher at 15.9±12.1%, indicating a significant difference ($p < 0.001$). A similar trend was observed when comparing surgical status; patients undergoing surgery exhibited significantly higher RLI values (17.3±12.5% vs. 4.4±10%, $p<0.001$).

Interestingly, 93 out of 465 (20%) patients had a negative RLI, suggesting that the unaffected kidney was larger than the affected one. A smaller proportion of these patients underwent surgery compared to those with positive RLI values (10.7% vs 32.0%, $p <$

0.001). A positive RLI increased the likelihood of surgery nearly fourfold (OR 3.9, 95% CI 2.0-7.7, p<0.001).

To gauge RLI's predictive ability for surgery, a univariable model was employed and proved significant (chi2(1) = 121.7, p<0.001), with a sensitivity of 48.1%, specificity of 95.5% and forecasted surgery correctly in 82.4% of predictions. Each unit increase in RLI correlated with a 13% increase in the odds of surgery (OR 1.13, 95% CI 1.10 – 1.16, p < 0.001). A cut-off value of 11.8% provided the optimal balance between sensitivity and specificity, as evidenced by a Youden J index of 0.514. At this cut-off point, the area under the ROC curve was 0.76, reflecting acceptable discriminative ability.

Model performance

RLI showed a great ability to differentiate between those requiring surgery to those resolving spontaneously, with area under the ROC curve of 0.804 (Figure 3) (bias-corrected c-statistics: 0.80, 95% CI 0.74-0.85). Visual assessment of the calibration plot reflects the model's ability to provide accurate probability estimates across the range of predicted risks (Figure 4). There is a slight underestimate of the risk of surgery for patients with an observed risk of <0.15 and for the ones whose observed risk of surgery is greater than 0.5. However, the Hosmer-Lemeshow goodness-of-fit test (chi2(8) = 20.99, p = 0.007) suggests a lack of fit between these observed and predicted values, and ultimately, potential room for improvement in model specification or inclusion of additional relevant variables.

Main Model with natural units

The initial model with 6 predictor variables - age, sex, hydronephrosis laterality, SFU score, APD, and the RLI - was statistically significant chi2(5)=184.02, p<0.001. Laterality, sex and age met the criteria for exclusion and thus were removed from the model (Table 4).

A backward elimination process identified three variables out of five potential predictors to be included in the final model: SFU, APD, and RLI (Table 5). This model included variables at their natural unit. The likelihood ratio test of the final model, compared to the initial model yielded a non-significant result (chi2(2) = 4.22, p=0.1212). This suggested that the model with fewer predictors was more appropriate. Furthermore, Akaike's Information Criterion (AIC) supported this, showing 8(df)191.2 for the initial model compared to 6(df)191.5 for the final model. The final logistic model was confirmed to be statistically significant with a chi2(5) = 186.57 and p < 0.001. Variables on the model did not show signs of multicollinearity (Table 6).

Both continuous variables, APD and RLI, satisfied the assumption of linearity, as indicated by the Box-Tidwell tests with p-values of 0.187 and 0.615, respectively. Consequently, there was no need for a restricted cubic spline for these variables. This was additionally confirmed by plotting the predicted log odds against both observed variable values and

observing a linear trend. An analysis of risk factors revealed that surgical risk escalated with increasing APD values (Figure 5). Specifically, patients with an APD of 20mm faced a surgery risk exceeding 30%. Likewise, RLI exhibited a linear relationship with surgical risk, incrementally rising as RLI values increased (Figure 6). The risk of surgery also noticeably increases with SFU (Figure 7), with a more than six-fold increase between SFU 2 and 3 and doubling from grade 3 to 4.

Model performance

The final model showcased excellent discriminative capabilities on the derivation cohort, with a c-statistic of 0.924 (Figure 8). After 1000 iterations of bootstrapping, the model continued to display excellent performance (bias-corrected c-statistics: 0.92, 95% CI 0.88 – 0.95). When applied to the validation cohort, the model maintained its discriminative ability (c-statistics: 0.94) (Figure 9). Calibration of the model was rigorously assessed through both graphical methods and the Hosmer-Lemeshow goodness-of-fit test, confirming an excellent fit (HL chi2(8) = 3.34, p = 0.911). The calibration plot revealed that the model's predictions were consistent and accurate across all the different ranges of observed risk (Figure 10). This calibration was maintained in the validation cohort (HL chi2(8) = 7.25, p= 0.510), displaying strong calibration with a slight underestimation of surgery for patients with intermediate-risk outcomes (Figure 11).

The HERO model: a clinical prediction model using categorical variables

To facilitate easier clinical application, the Hydronephrosis Early Risk of Obstruction (HERO) score model was introduced. This model employed categorized variables—APD, SFU, and RLI—from the final statistical model. Continuous variables were categorized by selecting points that optimized the Youden J index (thus, sensitivity and specificity) and rounded to the closest clinically significant values based on expert's consensus and cut-offs from previous literature [40] as mentioned in our methodology (Table 8).

In terms of categorical breakdowns, analyzing the distribution of these thresholds APD cut-offs could be found at the 25th, 50th, 75th, and 90th percentiles, while the RLI cut-offs were defined at the 25th, 45th, 60th, and 75th percentiles. SFU retained its original ordinal values. The distribution of these variables across categories is highlighted in Table 9. Distribution of patients undergoing surgery varied across different severity categories in the derivation cohort (Fisher's $p<0.001$ for all): SFU 4 was responsible for 92.7% (51) of surgical cases, RLI $\geq$ 15% for 68.9% (51) and APD $\geq$ 20 for 90% (36) (Table 10). There was no multicollinearity present on these variables, with a mean model VIF of 2 (Table *7*).

The HERO score emerged as a significant predictor of surgery, ($chi2(11) = 193.03$, $p<0.001$). The variables associated with the highest risk of surgery were SFU 4 (aOR 161.0; 95% CI 11.08-2,339.00), and APD ≥ 20mm (aOR 13.29; 95% CI 2.73-64.68) (Table 11). This model accurately classified 88.5% of surgical cases, achieving a sensitivity of 65.88% and a specificity of 96.93%. The area under the curve (AUC) was 0.935 for the derivation

cohort. Importantly, the HERO score did not deviate significantly from the model using natural units (LR chi2(6) = 6.46, p=0.373).

Model performance

The HERO score maintained excellent performance in distinguishing between patients requiring surgery (c-statistics: 0.93, bias-corrected c-statistics: 0.91, 95% CI 0.89-0.96) (Figure 12). When applied to the validation cohorts, the model maintained its ability to discriminate (c-statistics: 0.95) (Figure 13).

This model calibrated the data adequately derivation cohort (HL chi2(8) = 2.28, p = 0.971). The calibration plot analysis showed excellent calibration across the entire range of predicted probabilities (Figure 14). In the validation cohort, the model was well calibrated as per the Hosmer-Lemeshow test (HL chi2 (8)=6.05, p= 0.642), which was supported by evaluating the predicted-to-observed plot with a negligible underprediction for patients with risk of surgery between 0.4 and 0.6 (Figure 15).

Simplified HERO model: attributing scores

To increase usability in clinical daily practice, I simplified the HERO by assigning values to each HERO category based on the log odds from the logistic model and corresponding nomogram (Figure 16). This simplified version uses the same variables as the original HERO model. SFU highest categories 3 and 4 were assigned scores of 5 and 10 respectively, the double of their log-odds values. APD ≥ 20mm, present in 12.8% of

patients, was assigned a score of 5 and RLI ≥ 15%, present in 23.6% of the derivation cohort population, was assigned a score of 2. Scores for the remaining categories are informed in Table 12.

A score of 8 was the point that maximized the sensitivity (0.91) and specificity (0.79) of our model with great discrimination (AUC=0.85). A score equal to or over 8 was responsible for 91.8% of all surgeries on the derivation model (78/85). Scores up to 1.5 perfectly predicted non-surgical cases, with no cases undergoing surgery and almost all scores over 13 perfectly predicted surgery, with no cases foregoing surgery (Table 13,Table 14).

For easier interpretation, I have combined scores into groups of a similar range of risk of surgery: patients with scores from 0-6 points had the lowest risk of surgery at 8.7% risk of surgery, scores from 6.5 to 10.5 intermediate risk with surgical risk ranging from 17 to 40% and the highest risk of (92.5%) on patients scoring above 15 points (Table 15).

Model performance

The simplified model demonstrated good calibration in the derivation cohort (HL chi2(7)=2.91, p=0.893) (Figure 19) and exhibited the highest overestimation of surgical outcomes for patients with lower risk patients of surgery among all models (Figure 20). The calibration was excellent Despite this, the model still showed outstanding discrimination in both the derivation (c-statistics: 0.92, bias-corrected c-statistics: 0.91, 95% CI 0.87 - 0.95) (

Figure 17) and validation cohorts (c-statistics: 0.91, Figure 18).

Dichotomous HERO Model variables

To further enhance the HERO score's usability, I created a dichotomized model, aiming to evaluate its performance in predicting surgical outcomes. I aimed for an even more simple model that would have better performance in forecasting surgery. Cut-off points used for this model were defined by selecting the point that maximized Youden, i.e., the optimal threshold where these variables are most effective in correctly identifying true positives and true negatives. It yielded a model with 3 categories.

SFU 3 was the cut-off point with AUC=0.82 with sensitivity of 0.95 and specificity of 0.96. RLI optimal threshold was 12% with a sensitivity of 0.71, specificity of 0.80 and AUC of 0.76. Lastly, the APD cut-off point was 13mm, with a sensitivity of 0.71, specificity of 0.90 and AUC of 0.81. For patients undergoing surgery on the derivation cohort, 69.8% (60) had RLI ≥12%, 94.2% (81) were SFU 3 or 4, and 72.9% (62) presented an APD ≥ 13mm. The variable associated with the highest prediction of surgery was SFU 3-4 category with (aOR 9.8; 95% CI 3.45-27.68) (Table 16).

Model performance

The model dichotomizing the natural values of the 3 predictors had great performance. The dichotomized score showed excellent discrimination abilities in the derivation (c-statistics: 0.89, bias-corrected c-statistics: 0.84, 95% CI 0.78-0.89) (Figure 25) and validation cohorts (c-statistics: 0.91) (Figure 26), hence showing minimal change compared to the original main model. There was a minor drop in the discrimination, as expected since we miss the

granularity of information during the dichotomizing process. Notedly, 3% drop from the model on natural units and 4% from the HERO score.

The derivation cohort showed strong calibration across the entire range of predicted probabilities (Figure 27), confirmed by the Hosmer-Lemeshow test (HL chi2(6) = 3.68, p= 0.4513). In the validation cohort, the model was well calibrated and showed a small underprediction of risk or surgery for patients with observed risk around 30% (HL chi2(6) = 3.68, p= 0.4513) (Figure 28)

Simplified dichotomous score (dichotomous HERO)

A scoring system was also applied to the dichotomous model, with values based on the odds ratio from the regression. SFU grades 3-4 (present in 49.0% of patients from the derivation cohort) was scored 10 points, APD ≥ 13mm (present in 30.0% of patients from derivation cohort) was scored 5 points and RLI ≥ 12% (present in 32.5% of patients from derivation cohort) was scored 4 points. The score ranged from 0 with the absence of all categories to 19 with the presence of the 3 categories (Table 18).

Model performance

The simplified dichotomous model correctly classified 84.3% of patients undergoing surgery with a sensitivity of 72.9% and specificity of 88.6%. This model had great ability in discrimination between surgical patients and non-surgical outcomes in the derivation cohort (c-statistics:0.89, bias-corrected c-statistics:0.84, 95% CI 0.78-0.89) and validation cohort (C-statistics 0.91) (Figure 21, Figure 22). The model showed good calibration across

the different ranges of risk on derivation and validation cohorts (Figure 23, Figure 24). The surgical risk was 1.9% in patients with a score of 0, 5.4% in patients with a score of 4, 7.0% in patients with a score of 5, 22.5% in patients with a score of 10, 46.0% in patients with a score of 14, 52.8% in patients with a score of 15, 76.7% in patients with a score of 19 (Table 17).

HERO score vs Dichotomous HERO score

The HERO score was compared to the dichotomous score and showed a strong difference between models (Likelihood-ratio test chi2(8) = 43.46, p<0.0001). The dichotomous model, with a correct classification rate of 84.3%, had a sensitivity of 72.9% and a specificity of 88.6%. Additionally, its Positive Predictive Value (PPV) was recorded at 70.4%. Contrarily, the HERO score demonstrated an enhancement in specificity at 96.9%, with its sensitivity charting at 65.9%. This model's PPV stood at 88.9%, suggesting higher reliability in forecasting positive outcomes.

Sensitivity analyses

Unweighted simplified HERO score

The unweighted simplified HERO score was created by considering equal weights and attributing 1 point for each category. This simplified score varied from 0 in the absence of all variables, to 12 in the presence of all variables. Each variable ranged from 0 to 4 points.

On the derivation cohort, the unweighted simplified score varied from 1 to 12, with scores of 3 (13.42%, 42) and 6 (12.46%, 39) being the most frequent categories (Table 19).

The unweighted HERO model had 70.6% sensitivity and 93.4% specificity in the derivation cohort and forecasted surgery correctly in 87.22% of predictions. The unweighted model showcased excellent discriminative capabilities in both the derivation (c-statistics: 0.90, bias-corrected c-statistics: 0.88, 95% CI 0.83-0.92) and validation cohorts (c-statistics: 0.92) (Figure 29, Figure 30). A HERO score of 8 maximized the sensitivity (0.80) and specificity (0.85). HERO score equal to or greater than 8 is responsible for 80.0% (68) of surgical patients.

The model's calibration was evaluated using a graphical comparison and the Hosmer-Lemeshow goodness-of-fit test, which confirmed a good fit (chi2(9) = 5.39, p= 0.612). The calibration plot analysis showed strong calibration across the entire range of predicted probabilities for the derivation cohort (Figure 31). In the validation cohort, the model was calibrated (chi2(9) = 4.93, p= 0.669), yet the predicted-to-observed ratio suggested overprediction of surgical outcomes across patients with low to mid-risk of surgery (10-45%)(Figure 32).

Model ability to predict obstruction

The predictive capacity of the simplified model was evaluated for other outcomes of interest related to hydronephrosis. For this part of the sensitivity analysis, I assessed its forecasting abilities for the presence of an obstructive renal scan. Of the total of 314 patients

in the derivation cohort, 106 (33.8%) exhibited an obstructive renal scan. No significant difference was observed in the distribution of patients with obstructive renal scans between the derivation (106, 33.8%) and validation cohorts (56, 37.1%) (p=0.533). Patients with obstructions had double the mean APD value (17.9±8.6mm vs 8.0±5.1mm, p<0.001) and five times the RLI (3.3±9.8% vs 16.3±11.9%, p<0.001) compared to non-obstructed patients in the derivation cohort. A total of 83 (78.3%) of obstructed patients underwent pyeloplasty in derivation and 42 (97.7%) validation cohorts.

Simplified dichotomized model and obstruction

The simplified dichotomized model to forecast obstruction was significant (LR chi2(1) = 95.07, p<0.001). Variables indicating the highest risk of obstruction mirrored those in the surgery forecasting model. It correctly classified 85.3% of patients with a sensitivity of 81.9% and specificity of 87.02%. The score of 14 was the value that had the greatest accuracy with a sensitivity of 0.82 and specificity of 0.87, with an AUC of 0.84 and 46.8% risk of obstruction. Patients with a score up to 5 had a maximum of 11.9% chance of being obstructed. The highest risk of obstruction was seen in patients with a score of 19 with an 85% chance of being obstructed. Table 20 shows the risk of obstruction for different predictive scores.

The model maintained good calibration in the derivation cohort (HL chi2(4)=4.97, p=0.290). it underestimated the risk of obstruction by half in patients with a risk of obstruction of 50-60% in the validation cohort (Figure 33). Discrimination was exceptional

in both the derivation (c-statistics: 0.89, bias-corrected c-statistics: 0.85, 95% CI 0.79-0.90) and validation cohorts (c-statistics: 0.91).

Missing data

In the derivation cohort, age at baseline was missing for 2 entries, and APD for 1 entry. The validation cohort had no missing data. Given that our sample size resulted in less than 1% missing data, no multiple imputation method was applied. The dataset was utilized in its original format.

CHAPTER 4: DISCUSSION

Principal findings

Among a cohort of 465 patients with isolated unilateral UPJO-like hydronephrosis, 27.7% (129) underwent surgery. This study evaluated parameters on the baseline (initial postnatal) ultrasound showing rates of surgery and conservative management consistent with data available in the literature, which shows a 70% rate of spontaneous resolution (conservative management)[6,11]. Given that one in 500 live births screened on prenatal ultrasound is reported to have UPJO [3], this suggests 60 children for every 100,000 live births would likely undergo surgical intervention to resolve their hydronephrosis and can benefit from being counselled with a validated tool that informs a risk of surgery tailored to their individual measures from renal ultrasound. This thesis demonstrated that surgery can be effectively predicted using baseline ultrasound findings. The main predictor was SFU grading, particularly the SFU 4 category, most likely due to the presence of parenchymal thinning[27]. Parameters included in our model are easily identified and standardly reported on renal ultrasound as suggested by the Canadian Urological Association (CUA) and EAU for evaluation of severity and size of hydronephrosis during follow-up[2,29]. The present score – focused on baseline postnatal ultrasound - can enhance planning for closer follow-up for patients with higher risk and timely corrective measures whenever necessary before developing any potential irreversible renal damage. Concomitantly, potentially reducing

the burden of unnecessary examinations or follow-up for patients with low risk of surgery who will most likely resolve spontaneously[54].

Selecting the best model

The derivation final model containing the 3 predictors (APD, SFU, RLI) on its natural units predicted surgery with great discrimination and performance. However, the model was complex to implement in daily clinical practice and is better suited for a risk calculator. Thus, I created the HERO model categorizing these variables into four easily understandable categories to facilitate use. Despite the expected knowledge of losing granularity by categorizing variables[55], the model was not significantly different from the one with natural units and presented equivalent discrimination and calibration; therefore, information loss was not appreciated for the goal of increasing user-friendliness. A sensitivity analysis was performed to identify the discrimination and calibration between a weighted versus unweighted simplified model (i.e., categories within a variable increased by the rate of 1 point). A comparable ability performance was observed between them, thus an unweighted simplified model seemed to be the better option when prioritizing simplicity of use in clinical daily life.

The HERO model however still had a total of 12 categories, making it potentially cumbersome for clinicians to utilize and calculate. Thus, a dichotomized model was developed to test whether a reduction in categories would preserve the model's predictive

accuracy while simplifying its application. The dichotomized score was compared to the HERO model and showed a significant difference between them, confirming the small, yet potentially negligible, loss in clinical information. This dichotomized model showed only a marginal reduction in discrimination (a 4% absolute reduction in the AUC), an 8% decrease in identifying true negatives, and a 7% improvement in correctly identifying patients requiring surgical intervention.

When analyzing the odds ratio for the most severe category of SFU there was an unusually large OR of 160 with an extremely wide 95% CI (11.1-2,238.9) along with very high standard errors (SE) 219.8. This prompted me to evaluate the distribution of outcomes for this category. I noticed a quasi-complete separation for the SFU 1 category, i.e., it perfectly predicts failure, and there was a small cell count in the SFU 4 category for failure. The quasi-complete separation phenomenon leads to unstable and unreliable model estimates and can distort the performance metrics and usefulness to draw valid conclusions[56]. To address this issue, I performed an exploratory regression analysis using Firth's Method[56] - which penalizes the regression coefficient based on the large absolute values. However, the penalized OR for the SFU 4 category remained very high (OR 86.41, 95% CI 8.19-911.09). On the other hand, combining categories is another solution for addressing the phenomenon of small cell counts [56] and, the dichotomized model seamlessly addressed this issue. The OR for the SFU 3-4 category was 9.8 with a 95% CI of 3.45-27.7 with an SE of 5.2, showing that this model metric was more stable and reliable for interpretation.

Lastly, due to the presence of complete separation and small cell values, I cannot discard that the HERO model could be justified as having "optimistic" estimates, potential overfitting, and lack of generalizability to broader populations [56]. Specifically, complete separation indicates an artifact in the model's capability to correctly classify surgical cases, which could be clinically valuable in this setting where the early identification of obstruction and need for surgery is critical. Therefore, the simplified dichotomized model emerges as the preferred tool for clinical practice due to the balance of model robustness, small reduction in predictive accuracy and clinical usability.

Interpretation of the model

The variables incorporated into our simplified model have been discussed in prior literature as influential factors in determining the risk of surgery. This study corroborates that a Renal Length Index (RLI) greater than 12% is a predictive factor for surgery, showing acceptable discrimination with a value of 0.76[2,8,42,57].

This study revealed that SFU 3-4 ratings, driven by SFU 4, had the strongest influence on patient's risk of surgery. The SFU 4 category is characterized by parenchymal thinning, either segmental or diffuse, which indicates significant structural changes in the renal parenchyma[27]. Such changes point to a loss of functional tissue[44], compromised nephrons and, consequently impaired renal function and filtration efficacy[58]. If left untreated, these patients are at high risk of irreversible renal damage and eventually progressing to end-

stage renal disease[9,10]. In these cases, surgery is indicated to correct structural abnormalities and obstruction, thereby improving urinary drainage and mitigating further renal damage[44,59].

APD was the second most influential variable in predicting the risk of surgery in our model. This finding aligns with existing literature that has identified APD as a prognostic factor for loss of renal function[14]. Elevated APD values are indicative of a greater degree of renal pelvic distention, which in turn suggests a greater degree of dilation of the renal pelvis. This dilation can escalate intrarenal pressure within the collecting system, thereby placing the renal parenchyma at risk and ultimately surgery[60]. For example, a study evaluating 134 infants found APD was a significant predictor of surgery, particularly noting that the surgical group had a mean APD double that of the group managed conservatively (31.9±13.9mm vs 16.55±4.7mm)[61]. Similarly, Dias and colleagues identified an APD greater than 18 mm as having 100% sensitivity and 86% specificity for predicting surgery in patients with antenatal hydronephrosis [22]. Our study, however, proposed a smaller threshold for surgery at 13 mm. This can be attributed to the surgical team's leader philosophy advocating early surgical intervention to avoid continued nephron injury secondary to the prolonged exposure to an environment of constantly high intrarenal pelvic pressure demonstrated by experimental animal models[62,63].

Lastly, renal length index greater than 12% was found to be an independent predictor of surgery in our cohort. Utilizing a ratio enabled us to explore variations in kidney sizes by

treating each patient as their control, without having to account for age-related changes in renal length[64]. Patients who underwent surgical treatment had an affected renal length that was, on average, 10.5 mm (±4.4 mm) greater than those treated conservatively [65]. While the exact mechanisms underlying variations in renal length with different severities of hydronephrosis and duration of obstruction are not entirely understood in humans, animal models provide some insights. For instance, Yang and colleagues used a rabbit model to demonstrate that while declines in baseline GFR could be observed in both early and late stages of ureteral obstruction, renal length showed significant variance between these stages[32]; A significant reduction in initial GFR coupled with a minor increase in kidney size suggests that the obstruction is in its early phase, indicating better reserve function and potential for recovery. Conversely, a pronounced drop in initial GFR accompanied by substantial kidney enlargement indicates that the obstruction is in a later stage, characterized by diminished reserve function and a lack of potential for recovery[32].

Comparison to other models

Previous prognostic models have been developed to predict the course and outcomes of hydronephrosis. This is the first study developing and internally validating a model with SFU, RLI and APD to forecast surgery. Earlier work from the group led to the development of the Pyeloplasty Prediction Score (PPS)[40]. This score also utilized the same prognostic variables and indicated that children with a score of 8 or higher are eight times more likely to undergo surgery, with a likelihood ratio of 7.8. however, its methodology was not

envisioned as a prognostic model, but as a diagnostic one, where estimates output was the likelihood of having a surgical outcome given a score; therefore, its primary limitation lies in its inability to offer an estimated probability of outcome development based on individual risk[66].

The Hydronephrosis index (HI) was the first prognostic score, developed by Shapiro and colleagues[67]. It aimed to evaluate either hydronephrosis improvement or deterioration based on serial ultrasounds. It was reported as the percentage of the total renal area not affected by hydronephrosis. It was calculated by subtracting the difference of the dilated pelvis and calyces by the total renal area, divided by the total renal area. The goal of the index was to show the progression of HN and not to forecast a particular outcome. Exact thresholds and clinical applicability in forecasting obstruction or non-obstruction remain undefined[68].

On a related note, Rickard et al. proposed another predictive model based on the ratio of renal parenchyma to hydronephrosis area (PHAR). This model demonstrated good performance with an AUC of 0.81 for identifying pyeloplasty in patients with UPJO[69]. However, the practical application of PHAR measurements presents challenges. For example, it requires precise delineation of renal measurements and often involves additional staff, which can be time-consuming and resource-intensive. In contrast, our proposed model streamlines this process by using simplified and easily identifiable variables that conform to standard renal ultrasound procedures.

Blending different imaging components, the Hydronephrosis Severity Score (HSS)[70], is derived from parameters of both ultrasound (SFU and APD) and the differential renal function and drainage curve pattern to forecast surgical outcomes. This comprehensive approach has its merits but also comes with complications, such as its reliance on renal scan parameters that are susceptible to various influences (diuretic timing, hydration, and operator experience)[71]. Moreover, the HSS model does not uniformly use ultrasonographic measurements. Since SFU or APD have unequal ability to predict surgical outcomes, there is a possibility of a mismatch between these categories, leading to possible score discrepancies and an incomplete risk assessment. Lastly, the measure does not consider morphological abnormalities, such as the presence of an intrarenal pelvis. In this manner, patients with this condition can have their risk underestimated by this score.

More recently, Hodhod and colleagues introduced another prognostic model: the MAG-3 Suspected Obstructive Scoring system (SOS)[72]. The score utilized measures of the DRF, T ½ time and shape of the renogram curve. This model demonstrated remarkable discriminative capacity (0.97) for predicting surgery and renal obstruction. However, its accuracy could be compromised by the specifics of institutional protocols and the variable maturity of renal function in the studied population. Median age at presentation and baseline MAG-3 were 0.3 months (10 days) for the pyeloplasty and 0.6 months (18 days). Renal maturity happens at approximately 18 months of age, and the immature kidney has a lower glomerular filtration function and thus can lead to inaccurate assessment during the

calculation of renal function, and differential renal function during renal scan performance[73,74].

Strengths and Limitations

this study has multiple strengths. Primarily, it draws upon long-term data (15 years) from an academic, high-volume pediatric urology practice. This real-world dataset underwent rigorous prospective quality checks and auditing, resulting in minimal missing data (less than 1%). Additionally, the statistical methods employed for derivation, validation, and performance evaluation were robust, pre-specified, and in most cases, assessed using multiple approaches. By early detecting surgical candidates, we can potentially reduce the emotional and financial burden of disease associated with multiple follow-ups and imaging.

This thesis, however, is not exempt from limitations. The model was derived and internally validated using data from a single high-resource center and is pending external validation; thus, the generalizability and performance of the model scores might not represent accurately cohorts from other clinical practices or low-resource settings. We did not evaluate changes between successive clinical visits, thus we lack data on the magnitude and direction of this variability during clinical follow-up or its potential effect on the outcome of interest.

It's also important to note that assigning values to the simplified score based on odds ratios from logistic regression may have led to inconsistencies in value distribution on the HERO

model. This is particularly evident in the SFU 4 category, a quasi-complete separation with odds ratio. It is unclear, however, if this could have been mitigated by a larger sample size and potentially larger variability in this variable.

Additionally, the categorization of variables for ease of use in clinical practice may have resulted in underpowered categories for the HERO model, potentially masking the true effect size of these predictors, particularly the renal length index. Furthermore, during sensitivity analysis, the renal scan obstruction was defined per institutional urological department definition assessing T ½ time and characteristics of the drainage curve; this might not be representative of other institutions (hydration, bladder filling) or providers (urologist vs radiologists vs pediatric surgeons) definition of obstructive uropathy.
Lastly, our study did not include patients with different HN etiologies or the presence of bilateral disease. therefore, future studies must explore this score performance and identify outcomes of interest and patience with different hydronephrosis etiologies.

Additional considerations

Our patient analysis considered all subjects to have renal obstruction until proven otherwise. Given the surgical preferences and experience of the lead urologist, surgeries were conducted at early signs of obstruction rather than awaiting a decline in renal function. This could explain our lower cut-off points for surgical intervention. Moreover, the presence of an intra-renal pelvis was neither a criterion in our score nor information available in the dataset. Therefore, it is uncertain if patients with an intra-renal pelvis—

and consequently a smaller APD—may have been more likely to undergo surgery in our cohort.

Future Directions

Prognostic models like the present score serve as invaluable tools for seamlessly integrating structured risk assessments into patient counselling. They offer the dual benefits of enabling evidence-based, individualized care and efficiently allocating resources. By doing so, these models help to avoid unnecessary treatments and concentrate funds and timely interventions on those patients who need them most.

The logical next step for this research is the external validation of the dichotomized HERO score. The aim is to bolster the model's generalizability and robustness by employing real-world data from multiple centers. Furthermore, future studies should explore the model's usability and assess its long-term impact on clinical practice. One area of particular interest is the examination of how this proposed score fluctuates over time. Capturing such temporal variability could enhance the model's precision in clinical decision-making.

Another important avenue for future research is to explore variations in the interpretation and utility of the score among healthcare providers of different specialties. Understanding how general practitioners, pediatric surgeons, nephrologists, and urologists use and interpret this score could provide valuable insights. A prospective study that investigates the usability rates and outcomes across these specialties may provide clarity on this subject.

CHAPTER 5: CONCLUSION

The dichotomized HERO score, a prognostic model utilizing the following ultrasound-based measurements SFU 3-4, RLI >12%, and APD > 13mm, offers a precise method for identifying patients with isolated unilateral UPJO who are most likely to require surgical intervention. Demonstrating a 76.7% probability of surgery, 85% probability of renal scan obstruction when all three parameters are present and user-friendliness, this score holds considerable promise for clinical implementation. As a tool for individualized risk assessment, the score facilitates more targeted family counselling and has the potential to significantly reduce the emotional and psychological stress often associated with navigating complex medical landscapes. Furthermore, it offers healthcare providers a means for objective resource allocation, allowing for the deployment of resources and medical interventions to be prioritized based on the requirements of individual patient cases. This efficiency is particularly crucial in high-demand, resource-constrained settings - further enhancing its utility. As such, the score not only addresses financial considerations but also aligns seamlessly with the principles of patient-centred care.

References

1. Ficara A, Syngelaki A, Hammami A, Akolekar R, Nicolaides KH. Value of routine ultrasound examination at 35–37 weeks' gestation in diagnosis of fetal abnormalities. *Ultrasound in Obstetrics & Gynecology*. 2020;55(1):75-80. doi:10.1002/UOG.20857
2. Capolicchio JP, Braga LH, Szymanski KM. Canadian Urological Association/Pediatric Urologists of Canada guideline on the investigation and management of antenatally detected hydronephrosis. *Canadian Urological Association Journal*. 2018;12(4):85. doi:10.5489/CUAJ.5094
3. Liang CC, Cheng PJ, Lin CJ, Chen HW, Chao AS, Chang SD. Outcome of prenatally diagnosed fetal hydronephrosis. *J Reprod Med*. 2002;47(1):27-32. Accessed August 24, 2023. https://pubmed.ncbi.nlm.nih.gov/11838306/
4. Arena S, Chimenz R, Antonelli E, et al. A long-term follow-up in conservative management of unilateral ureteropelvic junction obstruction with poor drainage and good renal function. *Eur J Pediatr*. 2018;177(12):1761-1765. doi:10.1007/S00431-018-3239-2
5. ULMAN I, JAYANTHI VR, KOFF SA. The long-term followup of newborns with severe unilateral hydronephrosis initially treated nonoperatively. *J Urol*. 2000;164(3 Pt 2):1101-1105. doi:10.1097/00005392-200009020-00046
6. Koff SA, Campbell KD. The nonoperative management of unilateral neonatal hydronephrosis: natural history of poorly functioning kidneys. *J Urol*. 1994;152(2 Pt 2):593-595. doi:10.1016/S0022-5347(17)32658-7

7. Chertin B, Pollack A, Koulikov D, et al. Conservative treatment of ureteropelvic junction obstruction in children with antenatal diagnosis of hydronephrosis: lessons learned after 16 years of follow-up. *Eur Urol*. 2006;49(4):734-739. doi:10.1016/J.EURURO.2006.01.046

8. 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(3):266-271. doi:10.1016/J.JPUROL.2011.03.001

9. Calderon-Margalit R, Golan E, Twig G, et al. History of Childhood Kidney Disease and Risk of Adult End-Stage Renal Disease. *N Engl J Med*. 2018;378(5):428-438. doi:10.1056/NEJMOA1700993

10. Wühl E, van Stralen KJ, Verrina E, et al. Timing and outcome of renal replacement therapy in patients with congenital malformations of the kidney and urinary tract. *Clin J Am Soc Nephrol*. 2013;8(1):67-74. doi:10.2215/CJN.03310412

11. Passoni NM, Peters CA. Managing Ureteropelvic Junction Obstruction in the Young Infant. *Front Pediatr*. 2020;8:242. doi:10.3389/FPED.2020.00242/BIBTEX

12. 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(2):128-137. doi:10.1016/J.JPUROL.2019.01.011

13. Onen A. Grading of Hydronephrosis: An Ongoing Challenge. *Front Pediatr*. 2020;8:458. doi:10.3389/FPED.2020.00458/BIBTEX

14. Ucar AK, Kurugoglu S. Urinary Ultrasound and Other Imaging for Ureteropelvic Junction Type Hydronephrosis (UPJHN). *Front Pediatr*. 2020;8:562912. doi:10.3389/FPED.2020.00546/BIBTEX

15. Dy GW, Ellison JS, Fu BC, Holt SK, Gore JL, Merguerian PA. Variable Resource Utilization in the Prenatal and Postnatal Management of Isolated Hydronephrosis. *Urology*. 2017;108:155-160. doi:10.1016/J.UROLOGY.2017.05.042

16. Schaeffer AJ, Cartwright PC, Lau GA, et al. Utilization of Radiographic Imaging for Infant Hydronephrosis over the First 12 Months of Life. *Adv Urol*. 2020;2020. doi:10.1155/2020/2108362

17. Akhavan A, Shnorhavorian M, Garrison LP, Merguerian PA. Resource utilization and costs associated with the diagnostic evaluation of nonrefluxing primary hydronephrosis in infants. *J Urol*. 2014;192(3):919-924. doi:10.1016/J.JURO.2014.03.110

18. Buchanan CL, Morris MA, Matlock D, Kempe A, Vemulakonda VM. Parental experience and understanding of parent-provider discussions of treatment for infants with ureteropelvic junction obstruction. *PEC innovation*. 2023;2. doi:10.1016/J.PECINN.2023.100142

19. Timberlake MD, Anthony Herndon CD. Mild to moderate postnatal hydronephrosis--grading systems and management. *Nat Rev Urol*. 2013;10(11):649-656. doi:10.1038/NRUROL.2013.172

20. Moghazi S, Jones E, Schroepple J, et al. Correlation of renal histopathology with sonographic findings. *Kidney Int*. 2005;67(4):1515-1520. doi:10.1111/J.1523-1755.2005.00230.X

21. Nguyen HT, Benson CB, Bromley B, et al. Multidisciplinary consensus on the classification of prenatal and postnatal urinary tract dilation (UTD classification system). *J Pediatr Urol*. 2014;10(6):982-998. doi:10.1016/J.JPUROL.2014.10.002

22. Dias CS, Silva JMP, Pereira AK, et al. Diagnostic accuracy of renal pelvic dilatation for detecting surgically managed ureteropelvic junction obstruction. *J Urol*. 2013;190(2):661-666. doi:10.1016/J.JURO.2013.02.014

23. Sharifian M, Esfandiar N, Mohkam M, Dalirani R, Baban Taher E, Akhlaghi A. Diagnostic accuracy of renal pelvic dilatation in determining outcome of congenital hydronephrosis. *Iran J Kidney Dis*. 2014;8(1):26-30.

24. Burgu B, Aydogdu O, Soygur T, Baker L, Snodgrass W, Wilcox D. When is it necessary to perform nuclear renogram in patients with a unilateral neonatal hydronephrosis? *World J Urol*. 2012;30(3):347-352. doi:10.1007/S00345-011-0744-6

25. Hodhod A, Eid H, Capolicchio JP, et al. How can we measure the renal pelvic anteroposterior diameter in postnatal isolated hydronephrosis? *J Pediatr Urol*. 2023;19(1):75-82. doi:10.1016/J.JPUROL.2022.08.007

26. Rickard M, Braga LH, Oliveria JP, Romao R, Demaria J, Lorenzo AJ. Percent improvement in renal pelvis antero-posterior diameter (PI-APD): Prospective validation and further exploration of cut-off values that predict success after

pediatric pyeloplasty supporting safe monitoring with ultrasound alone. *J Pediatr Urol*. 2016;12(4):228.e1-228.e6. doi:10.1016/J.JPUROL.2016.04.003

27. Fernbach SK, Maizels M, Conway JJ. Ultrasound grading of hydronephrosis: Introduction to the system used by the society for fetal urology. *Pediatr Radiol*. 1993;23(6):478-480. doi:10.1007/BF02012459/METRICS

28. Keays MA, Guerra LA, Mihill J, et al. Reliability assessment of Society for Fetal Urology ultrasound grading system for hydronephrosis. *J Urol*. 2008;180(4 Suppl):1680-1683. doi:10.1016/J.JURO.2008.03.107

29. C. Radmayr, G. Bogaert, B. Burgu, et al. EAU Guidelines on Paediatric Urology. Accessed April 9, 2023. https://uroweb.org/guidelines/paediatric-urology/chapter/the-guideline

30. Braga LH, McGrath M, Farrokhyar F, Jegatheeswaran K, Lorenzo AJ. 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(3 Pt 2):831-837. doi:10.1016/J.JURO.2016.08.099

31. Sibai H, Pippi Salle JL, Houle AM, Lambert R. Hydronephrosis with diffuse or segmental cortical thinning: impact on renal function. *J Urol*. 2001;165(6 Pt 2):2293-2295. doi:10.1016/S0022-5347(05)66187-3

32. Yang Q, Wang C, Gao C, et al. Does baseline renal function always decrease after unilateral ureteral severe obstruction? -experimental validation and novel findings by Tc-99m diethylene triamine pentaacetate acid (DTPA) dynamic renal

scintigraphy. *Quant Imaging Med Surg*. 2019;9(8):1451-1465. doi:10.21037/QIMS.2019.07.09

33. Khazaei MR, Mackie F, Rosenberg AR, Kainer G. Renal length discrepancy by ultrasound is a reliable predictor of an abnormal DMSA scan in children. *Pediatr Nephrol*. 2008;23(1):99-105. doi:10.1007/S00467-007-0637-5

34. SPU - RELATIVE RENAL LENGTH DISCREPANCY AS AN INDICATOR OF RENAL OBSTRUCTION IN CHILDREN WITH UPJO-LIKE. Accessed May 28, 2023. https://fallcongress.spuonline.org/abstracts/2019/P2.cgi

35. Rongviriyapanich C, Sakunchit T, Sudla C, Mungkung S, Pongnapang N, Yeong CH. Sonographic renal length and volume of normal Thai children versus their Chinese and Western counterparts. *Clin Exp Pediatr*. 2020;63(12):491-498. doi:10.3345/CEP.2019.01676

36. Miletic D, ˇ eljko Fuč kar Z, ˇ ustic AS, Mozetič V, ˇ timac DS, ˇ auhar GZ. Sonographic Measurement of Absolute and Relative Renal Length in Adults. *J Clin Ultrasound*. 1998;26:185-189. doi:10.1002/(SICI)1097-0096(199805)26:4

37. Currarino G, Williams B, Dana K. Kidney length correlated with age: normal values in children. *https://doi-org.libaccess.lib.mcmaster.ca/101148/radiology15036695070*. 1984;150(3):703-704. doi:10.1148/RADIOLOGY.150.3.6695070

38. Chen JJ, Pugach J, Patel M, Luisiri A, Steinhardt GF. The Renal Length Nomogram: Multivariable Approach. *J Urol*. 2002;168(5):2149-2152. doi:10.1016/S0022-5347(05)64339-X

39. Lane PH, Belsha CW, Plummer J, Steinhardt GF, Lynch RE, Wood EG. Relationship of renal size, body size, and blood pressure in children. *Pediatr Nephrol*. 1998;12(1):35-39. doi:10.1007/S004670050399

40. Li B, McGrath M, Farrokhyar F, Braga LH. Ultrasound-Based Scoring System for Indication of Pyeloplasty in Patients With UPJO-Like Hydronephrosis. *Front Pediatr*. 2020;8. doi:10.3389/FPED.2020.00353

41. Arora S, Yadav P, Kumar M, et al. Predictors for the need of surgery in antenatally detected hydronephrosis due to UPJ obstruction--a prospective multivariate analysis. *J Pediatr Urol*. 2015;11(5):248.e1-248.e5. doi:10.1016/J.JPUROL.2015.02.008

42. Longpre M, Nguan A, MacNeily AE, Afshar K. Prediction of the outcome of antenatally diagnosed hydronephrosis: a multivariable analysis. *J Pediatr Urol*. 2012;8(2):135-139. doi:10.1016/J.JPUROL.2011.05.013

43. Dhillon HK. Prenatally diagnosed hydronephrosis: the Great Ormond Street experience. *Br J Urol*. 1998;81 Suppl 2(SUPPL. 2):39-44. doi:10.1046/J.1464-410X.1998.0810S2039.X

44. Nguyen HT, Herndon CDA, Cooper C, et al. The Society for Fetal Urology consensus statement on the evaluation and management of antenatal hydronephrosis. *J Pediatr Urol*. 2010;6(3):212-231. doi:10.1016/J.JPUROL.2010.02.205

45. Grignon A, Filion R, Filiatrault D, et al. Urinary tract dilatation in utero: classification and clinical applications. *Radiology*. 1986;160(3):645-647. doi:10.1148/RADIOLOGY.160.3.3526402

46. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338(7708):1432-1435. doi:10.1136/BMJ.B605

47. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test*. 2018;2(3):249-262. doi:10.1007/S41664-018-0068-2/FIGURES/9

48. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365. doi:10.1371/JOURNAL.PONE.0224365

49. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379. doi:10.1016/S0895-4356(96)00236-3

50. Harrell FE. Multivariable Modeling Strategies. Published online 2015:63-102. doi:10.1007/978-3-319-19425-7_4

51. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0B013E3181C30FB2

52. O'Reilly P, Aurell M, Britton K, Kletter K, Rosenthal L, Testa T. Consensus on diuresis renography for investigating the dilated upper urinary tract. Radionuclides in Nephrourology Group. Consensus Committee on Diuresis Renography. *J Nucl Med*. 1996;37(11):1872-1876.

53. O'REILLY PH, TESTA HJ, LAWSON RS, FARRAR DJ, EDWARDS EC. Diuresis renography in equivocal urinary tract obstruction. *Br J Urol*. 1978;50(2):76-80. doi:10.1111/J.1464-410X.1978.TB03030.X

54. Thomas DFM. Prenatal diagnosis: what do we know of long-term outcomes? *J Pediatr Urol*. 2010;6(3):204-211. doi:10.1016/J.JPUROL.2010.01.013

55. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080. doi:10.1136/BMJ.332.7549.1080

56. Clark RG, Blanchard W, Hui FKC, Tian R, Woods H. Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data. *Research Methods in Applied Linguistics*. 2023;2(1):100044. doi:10.1016/J.RMAL.2023.100044

57. Coplen DE, Austin PF, Yan Y, Blanco VM, Dicke JM. The Magnitude of Fetal Renal Pelvic Dilatation can Identify Obstructive Postnatal Hydronephrosis, and Direct Postnatal Evaluation and Management. *J Urol*. 2006;176(2):724-727. doi:10.1016/J.JURO.2006.03.079

58. Konda R, Sakai K, Ota S, Abe Y, Hatakeyama T, Orikasa S. Ultrasound Grade of Hydronephrosis and Severity of Renal Cortical Damage on 99mTechnetium Dimercaptosuccinic Acid Renal Scan in Infants With Unilateral Hydronephrosis During Followup and After Pyeloplasty. *J Urol*. 2002;167(5):2159-2163. doi:10.1016/S0022-5347(05)65118-X

59. Onen A. Treatment and outcome of prenatally detected newborn hydronephrosis. *J Pediatr Urol*. 2007;3(6):469-476. doi:10.1016/J.JPUROL.2007.05.002

60. Sinha A, Bagga A, Krishna A, et al. Revised guidelines on management of antenatal hydronephrosis. *Indian Pediatr*. 2013;50(2):215-231. doi:10.1007/s13312-013-0064-6

61. Mahmoud TA, Morsy EE din S, Morsy HAEA, Abouzeid AM, Elmoghazy HM. Predictors of surgical intervention for antenatally detected ureteropelvic junction obstruction (UPJO): A prospective multivariate analysis. *Urologia*. Published online 2023. doi:10.1177/03915603231173009

62. Thornhill BA, Forbes MS, Marcinko ES, Chevalier RL. Glomerulotubular disconnection in neonatal mice after relief of partial ureteral obstruction. *Kidney Int*. 2007;72(9):1103-1112. doi:10.1038/SJ.KI.5002512

63. Chevalier RL. CAKUT: A Pediatric and Evolutionary Perspective on the Leading Cause of CKD in Childhood. *Pediatr Rep*. 2023;15(1):143. doi:10.3390/PEDIATRIC15010012

64. Kadioglu A. Renal measurements, including length, parenchymal thickness, and medullary pyramid thickness, in healthy children: what are the normative ultrasound values? *AJR Am J Roentgenol*. 2010;194(2):509-515. doi:10.2214/AJR.09.2986

65. Kelley JC, White JT, Goetz JT, Romero E, Leslie JA, Prieto JC. Sonographic Renal Parenchymal Measurements for the Evaluation and Management of Ureteropelvic Junction Obstruction in Children. *Front Pediatr*. 2016;4(MAY). doi:10.3389/FPED.2016.00042

66. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis*. 2013;11(SUPPL.1):129-141. doi:10.1111/JTH.12262

67. Shapiro SR, Wahl EF, Silberstein MJ, Steinhardt G. Hydronephrosis index: a new method to track patients with hydronephrosis quantitatively. *Urology*. 2008;72(3):536-538. doi:10.1016/J.UROLOGY.2008.02.007

68. Rud O, Moersler J, Peter J, et al. Prospective evaluation of interobserver variability of the hydronephrosis index and the renal resistive index as sonographic examination methods for the evaluation of acute hydronephrosis. *BJU Int*. 2012;110(8 Pt B):E350-E356. doi:10.1111/J.1464-410X.2012.11087.X

69. Rickard M, Lorenzo AJ, Braga LH. Renal Parenchyma to Hydronephrosis Area Ratio (PHAR) as a Predictor of Future Surgical Intervention for Infants With High-grade Prenatal Hydronephrosis. *Urology*. 2017;101:85-89. doi:10.1016/J.UROLOGY.2016.09.029

70. Babu R, Venkatachalapathy E, Sai V. Hydronephrosis severity score: an objective assessment of hydronephrosis severity in children-a preliminary report. *J Pediatr Urol*. 2019;15(1):68.e1-68.e6. doi:10.1016/J.JPUROL.2018.09.020

71. Wong DC, Rossleigh MA, Farnsworth RH. Diuretic renography with the addition of quantitative gravity-assisted drainage in infants and children. *J Nucl Med*. 2000;41(6):1030-1036.

72. Hodhod A, Fermin-Risso C, Farhad M, et al. Can we improve the usefulness of the diuretic renogram in the diagnosis of ureteropelvic junction obstruction (UPJO) in

children? Introduction of mercaptoacetyltriglycine-suspected obstruction scoring system (MAG-SOS). *J Pediatr Urol*. 2023;19(3):311.e1-311.e8. doi:10.1016/J.JPUROL.2023.02.022

73. Gordon I, Piepsz A, Sixt R. Guidelines for standard and diuretic renogram in children. *Eur J Nucl Med Mol Imaging*. 2011;38(6):1175-1188. doi:10.1007/S00259-011-1811-3

74. Caglar M, Gedik GK, Karabulut E. Differential renal function estimation by dynamic renal scintigraphy: influence of background definition and radiopharmaceutical. *Nucl Med Commun*. 2008;29(11):1002-1005. doi:10.1097/MNM.0B013E32830978AF

CHAPTER 6: APPENDIX

Table 1 Variables used in the initial logistic regression model

Variable	Type	Result
Anteroposterior pelvic diameter (APD)	continuous	mm
Age at baseline ultrasound	continuous	months
Society for Fetal Urology hydronephrosis grading system (SFU)	ordinal	I II III IV
Renal Length Index	continuous	%
Laterality	categorical	Right Left
Sex	categorical	Female Male

Figure 1- Patient flow chart

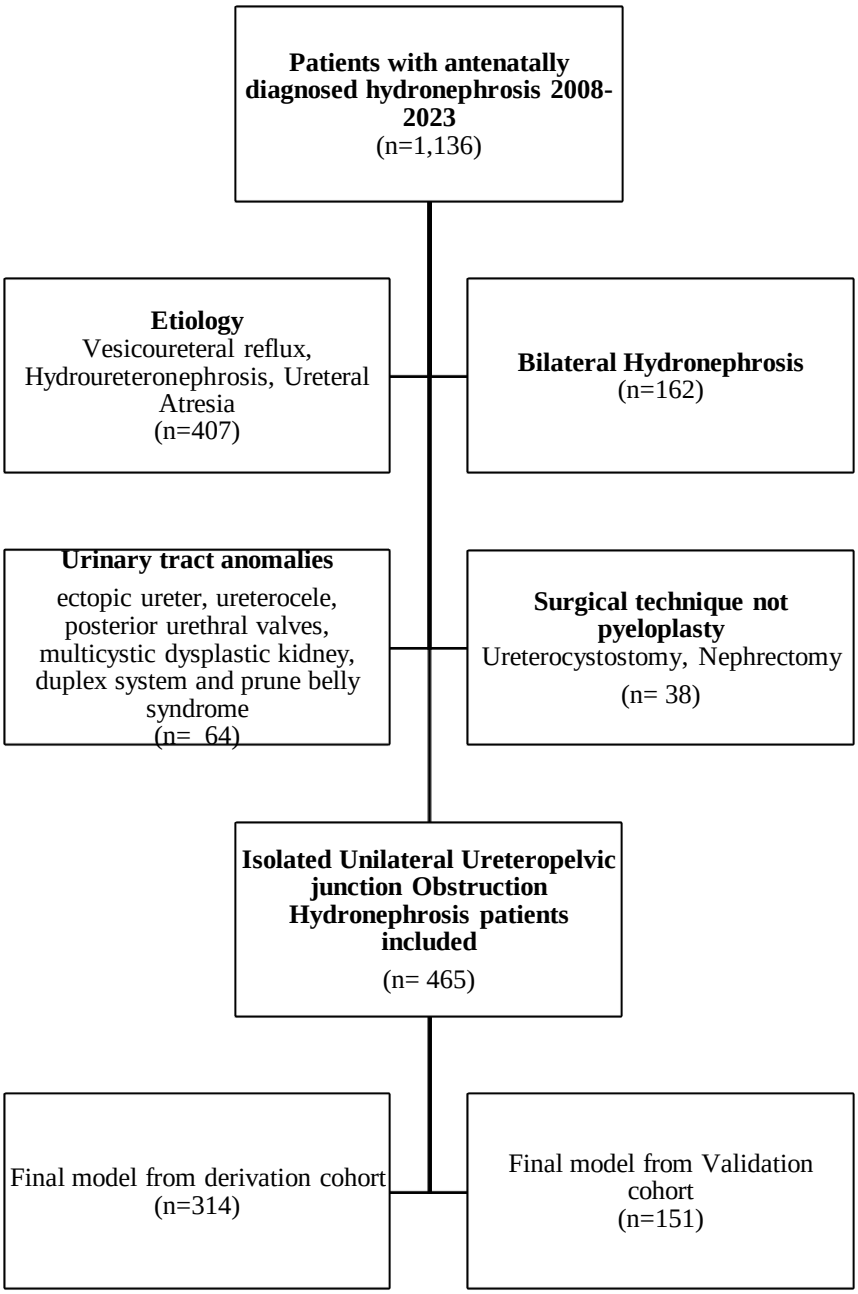


Table 2 - Patient characteristics

	All patients	Derivation Cohort	Validation Cohort	p-value
	N=465	N=314	N=151	
Age at baseline, m	3.8±3.8	3.7 ±3.9	4.0 ±3.6	0.47
Male	354 (76.1%)	243 (77.4%)	111 (73.5%)	0.36
Female	111 (23.9%)	71 (22.6%)	40 (26.5%)	
Laterality				
Left	355 (76.3%)	238 (75.8%)	117 (77.5%)	0.69
Right	110 (23.7%)	76 (24.2%)	34 (22.5%)	
SFU				
1	85 (18.3%)	56 (17.8%)	29 (19.2%)	0.98
2	154 (33.1%)	104 (33.1%)	50 (33.1%)	
3	144 (31.0%)	99 (31.5%)	45 (29.8%)	
4	82 (17.6%)	55 (17.5%)	27 (17.9%)	
APD, mm	11.2 ±7.8	11.3 ±8.0	10.9 ±7.3	0.58
RLI, %	7.0 (1.5-14.9)	6.8 (0-14.3)	8.1 (1.7-16.2)	0.24
Renal length, mm				
Affected	6.0 (5.5-6.6)	5.9 (5.4-6.6)	5.9 (5.5-6.7)	0.31
Contralateral	5.5 (5-6)	5.5 (5-6)	5.5 (5.0-5.9)	0.95
Renal scan parameters				
T ½ time, min				
< 30	385 (82.8%)	260 (82.8%)	125 (82.8%)	0.99
≥ 30	80 (17.2%)	54 (17.2%)	26 (17.2%)	
Drainage curve				
Non-obstructed	306 (65.8%)	211 (67.2%)	95 (62.9%)	0.36
Obstructed	159 (34.2%)	103 (32.8%)	56 (37.1%)	
Renal Scan decision				
Non-obstructed	303 (65.2%)	208 (66.2%)	95 (62.9%)	0.48
obstructed	162 (34.8%)	106 (33.8%)	56 (37.1%)	
Surgery				
No	336 (72.3%)	228 (72.6%)	108 (71.5%)	0.81
Yes	129 (27.7%)	86 (27.4%)	43 (28.5%)	
Months to surgery	10.9 ±17.5	11.9 ±19.7	8.9 ±12.0	0.36

Data are presented as n (%), median (IQR), mean ±standard deviation; SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length index; T ½ time: time in which half of isotope is drained.

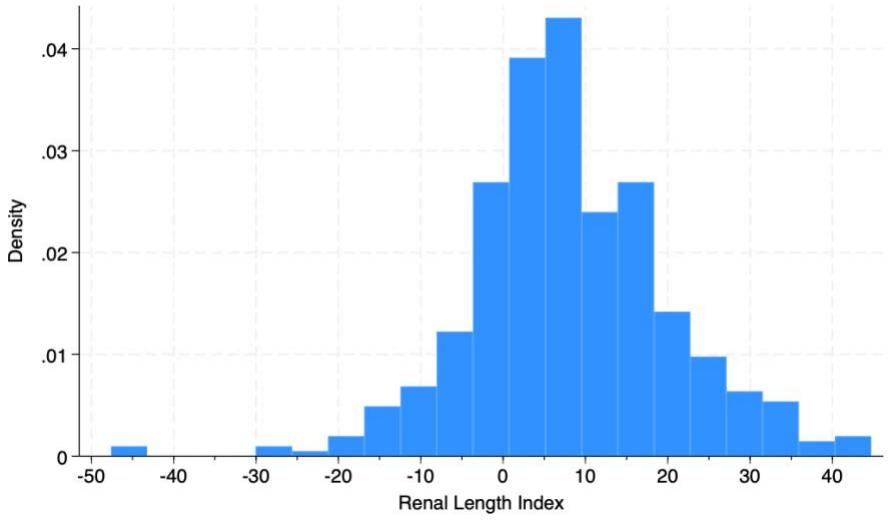
*t-test for Age at baseline, APD, months to surgery; Wilcoxon rank sum for RLI, renal length affected, renal length contralateral; Fisher exact test for all other variables

Table 3 - Patient characteristics by surgery outcome

	Surgery No	Surgery Yes	p-value
	N=336	N=129	
Age at baseline, month	2.9 (1.6-5.0)	1.8 (0.9-4.6)	0.84
Male	268 (79.8%)	86 (66.7%)	0.005
Female	68 (20.2%)	43 (33.3%)	
Laterality			
Left	270 (80.4%)	85 (65.9%)	0.001
Right	66 (19.6%)	44 (34.1%)	
SFU			
1	84 (25.0%)	1 (0.8%)	<0.001
2	149 (44.3%)	5 (3.9%)	
3	97 (28.9%)	47 (36.4%)	
4	6 (1.8%)	76 (58.9%)	
APD, mm	8.1±4.2	19.2±9.2	<0.001
RLI, %	5 (0-10.3)	18.3 (8.1-25.0)	<0.001
Renal length, mm			
Affected	5.8 (5.4-6.3)	6.6 (5.9-7.3)	<0.001
Contralateral	5.5 (5.0-6.0)	5.4 (5.0-5.9)	0.17
Renal scan parameters			
T ½ time, min			
< 30	325 (96.7%)	60 (46.5%)	<0.001
≥ 30	11 (3.3%)	69 (53.5%)	
Drainage curve			
Non-obstructed	302 (89.9%)	4 (3.1%)	<0.001
Obstructed	34 (10.1%)	125 (96.9%)	
Renal Scan decision			
Non-obstructed	299 (89.0%)	4 (3.1%)	<0.001
Obstructed	37 (11.0%)	125 (96.9%)	

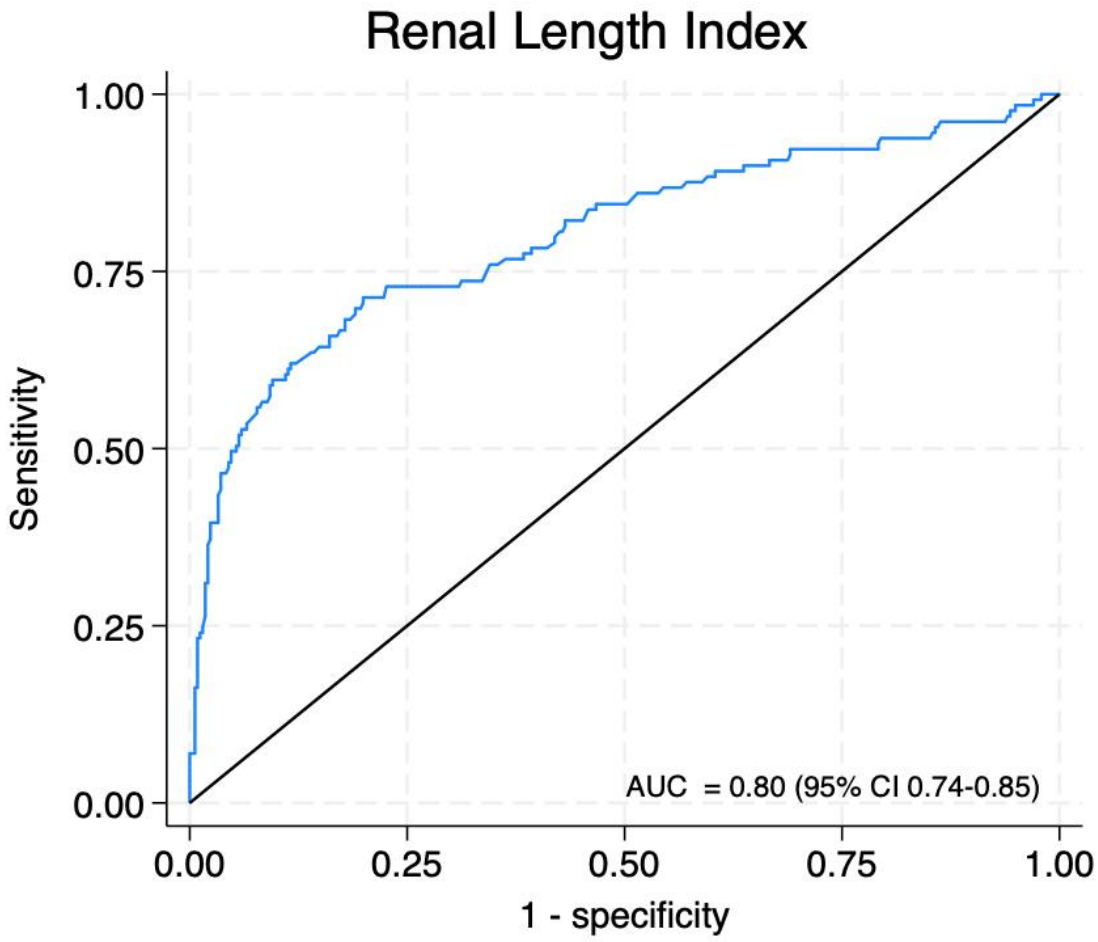
Data are presented as n (%), median (IQR), mean ±standard deviation; SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length index. T ½ time: time in which half of the isotope is drained.
*t-test for APD; Wilcoxon rank sum for Age at baseline, RLI, renal length affected, renal length contralateral; Fisher exact test for all other variables

Figure 2 - Renal Length Index distribution



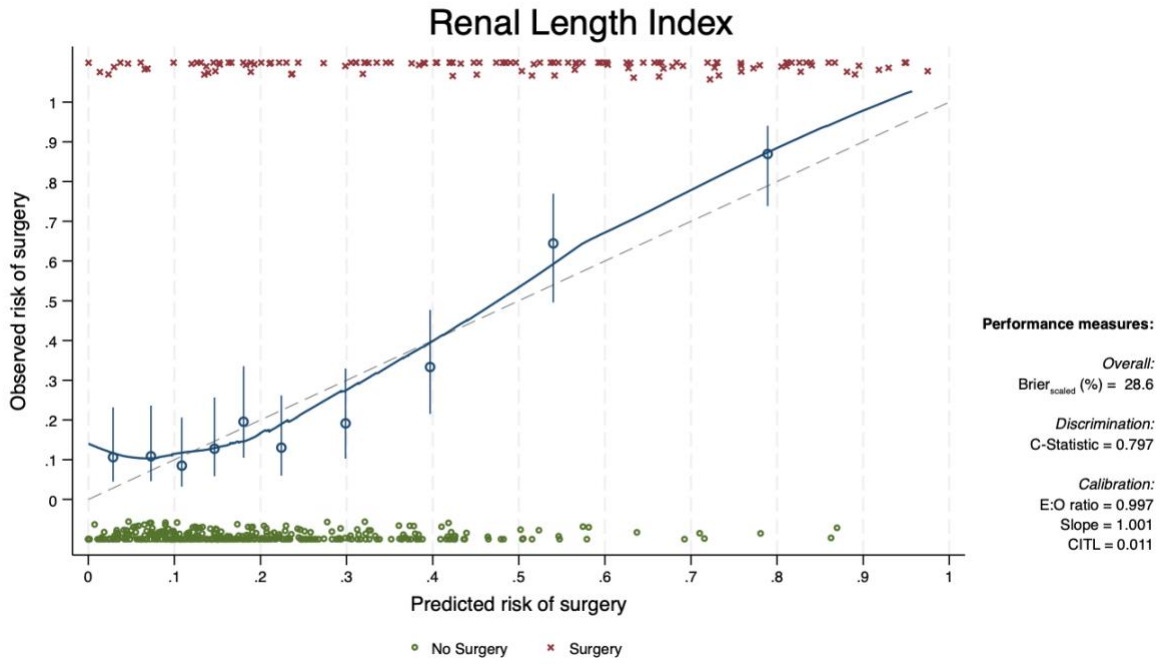
Histogram showing the normal distribution of the renal length index.

Figure 3- Discrimination of Renal Length Index for surgery



Receiver-operating characteristic curve indicating the good discrimination ability of Renal Length Index in predicting surgery. AUC: Area under the curve.

Figure 4- Calibration Plot of Renal Length Index



Calibration plot of the Renal Length Index. Y-axis: Observed risk of surgery. X-axis: Predicted risk of surgery. Reference (dashed): line representing perfect prediction risk equal to observed proportion of patients. Lowess (continuous, blue): non-parametric curve representing the plotted observed vs predicted outcomes.

Table 4- Derivation initial model

Independent variable (n=313)	Adjusted OR	95% CI	p-value
SFU			
1	Reference		
2	1.49	0.16-14.02	0.729
3	7.85	0.92-66.75	0.064
4	90.68	7.58-1,084.62	<0.001
APD, mm	1.15	1.05-1.25	0.001
RLI, %	1.02	0.97- 1.07	0.45
Laterality			
Right	Reference		
Left	0.93	0.38-2.28	0.866
Age at baseline, m	1.08	0.99-1.18	0.396

Multivariable logistic regression using full model. SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length index.
AIC: (8df) 191.24

Table 5 - Derivation of final model with natural units

Independent variable (n=313)	Adjusted OR	95% CI	p-value
SFU			
1	Reference		
2	1.49	0.16-13.86	0.728
3	7.57	0.89-64.13	0.063
4	78.57	6.65-928.02	0.001
APD, mm	1.15	1.05-1.25	0.001
RLI, %	1.02	0.97-1.06	0.474

Multivariable logistic regression final model. SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length Index. AIC: (6df) 191.5

Table 6 - Variance inflation factor of predictors included in final derivation model

Variable	VIF
APD	2.1
RLI	1.55
SFU	2.23
Mean VIF	1.96

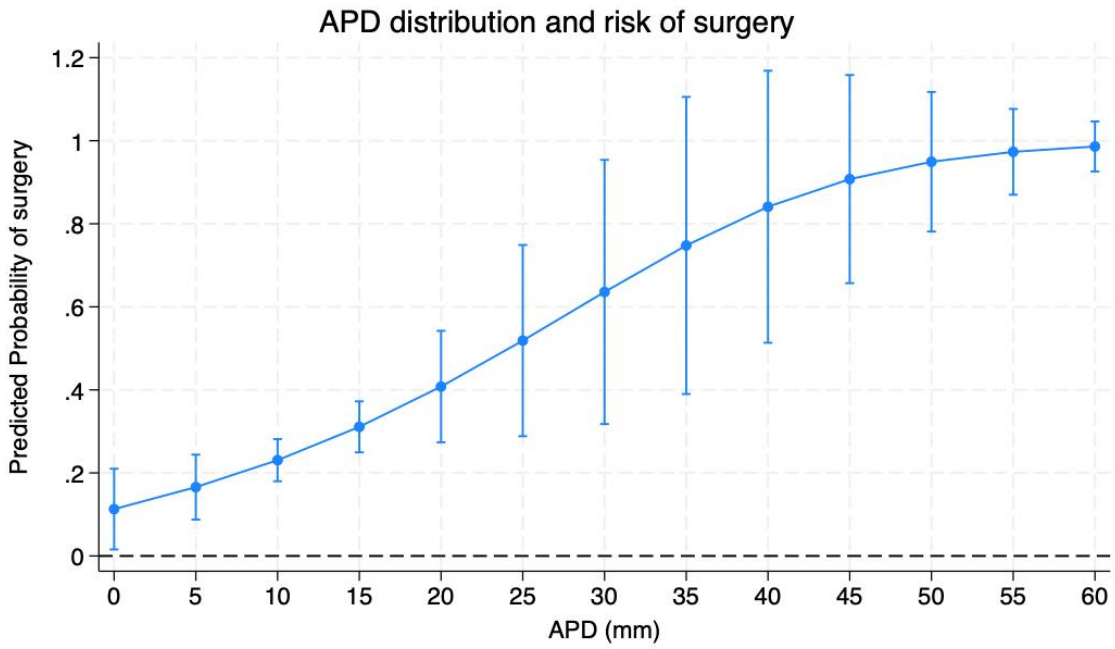
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length Index; VIF: variance inflation factor

Table 7- Variance inflation factor of HERO predictors (Derivation)

Variable	VIF
APD	
5-10mm	2.65
10-15mm	3.04
15-20mm	2.21
≥ 20mm	3.41
RLI	
1-5%	1.38
5-10%	1.56
10-15%	1.54
≥15%	2.41
SFU	2.23
2	2.18
3	2.98
4	3.78
Mean VIF	2.47

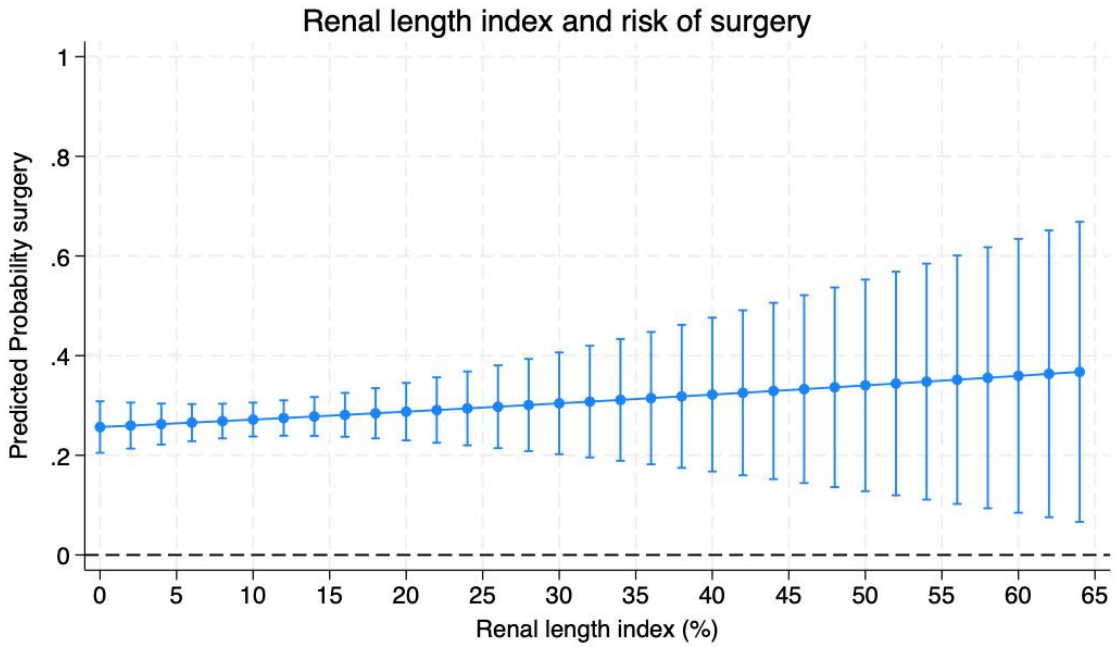
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length Index; VIF: variance inflation factor; HERO: Hydronephrosis Early Risk of Obstruction

Figure 5 - Risk of surgery by APD



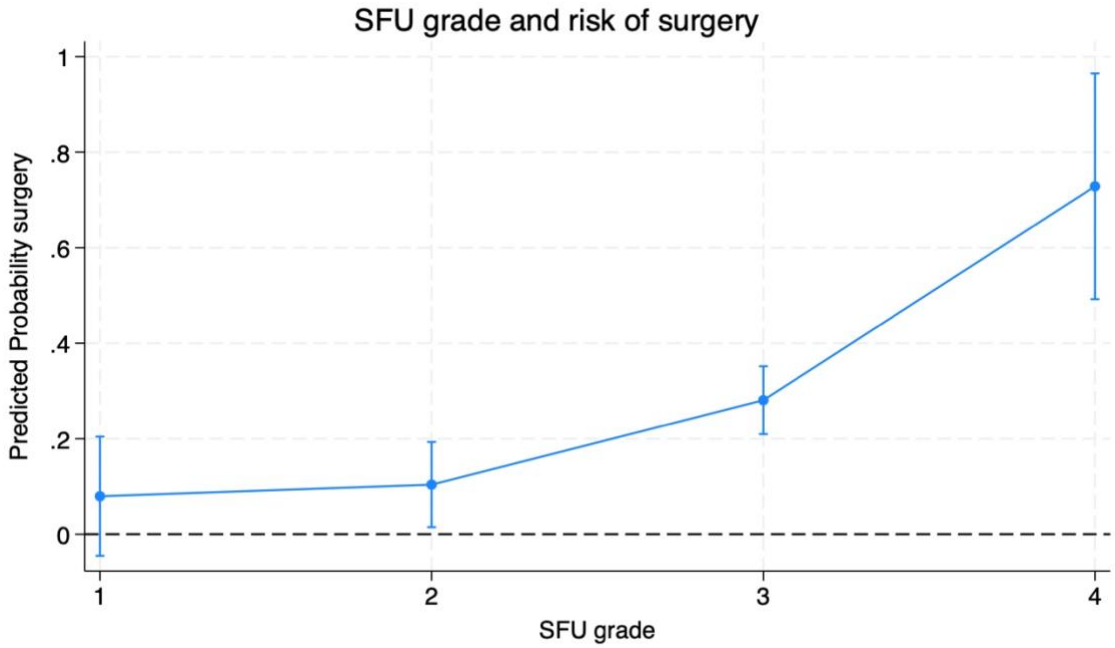
Graph illustrating the relationship between Anteroposterior pelvic (APD) and predicted risk of surgery. Y axis: predicted probability of surgery using the main model. X-axis: APD in 5mm intervals.

Figure 6 - Risk of surgery by RLI



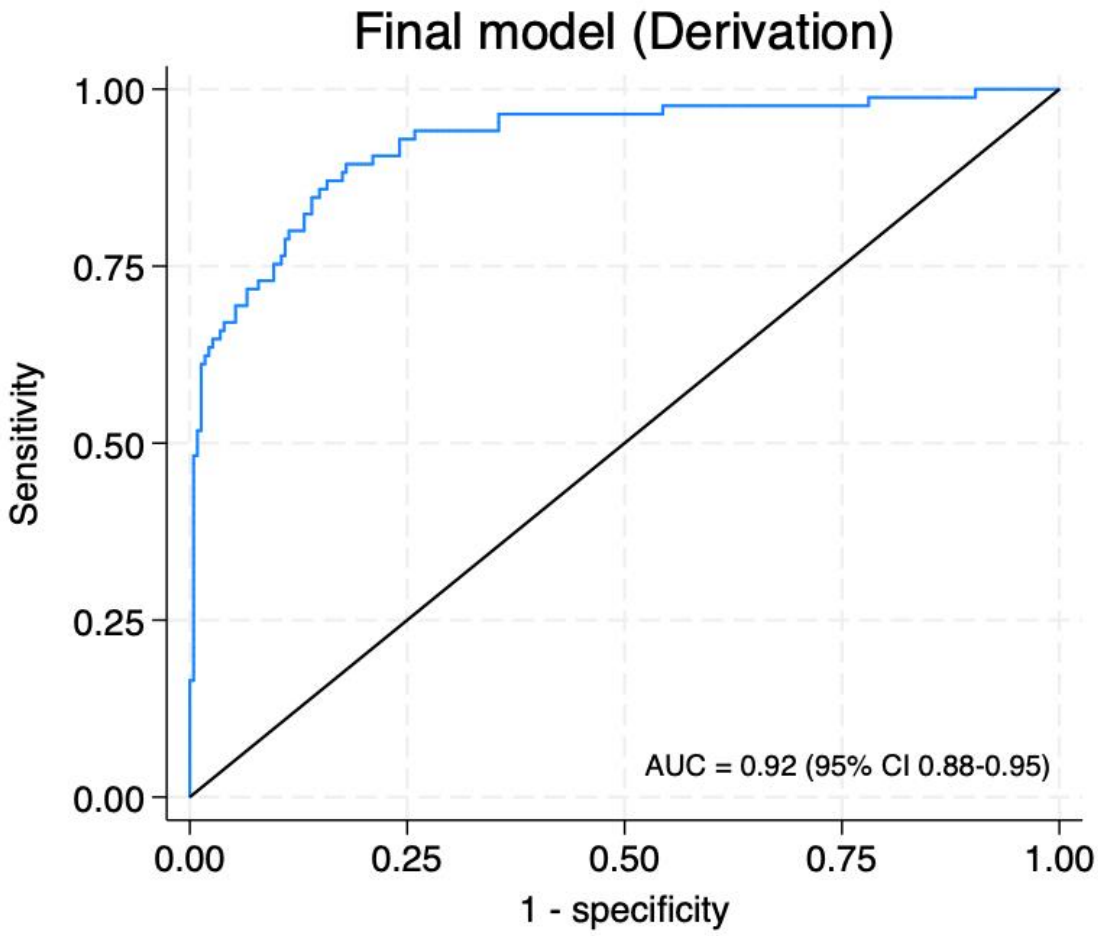
Graph illustrating the relationship between renal length index (RLI) and predicted risk of surgery. Y axis: predicted probability of surgery using the main model. X-axis: RLI in 5% intervals.

Figure 7 - Risk of surgery by SFU grade



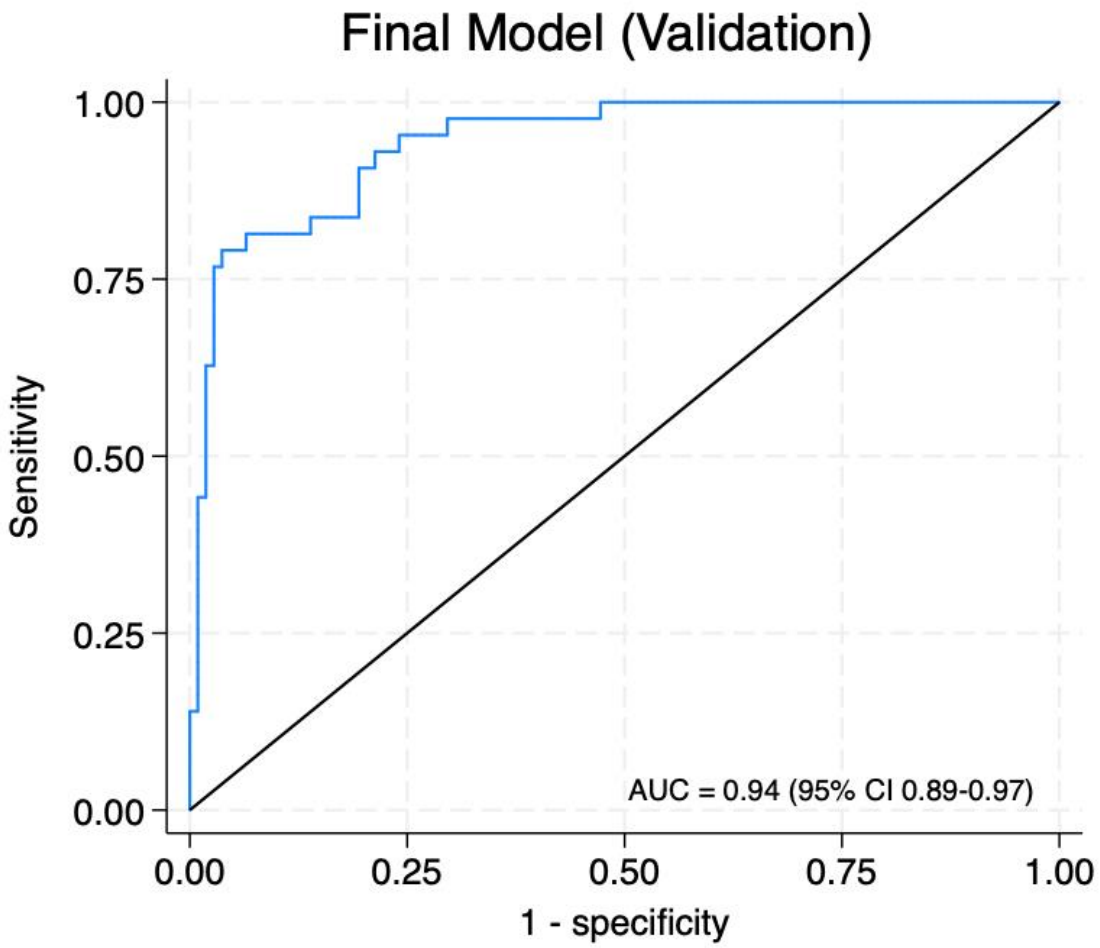
Graph illustrating the relationship between Society for Fetal Urology grade (SFU) and predicted risk of surgery. Y axis: predicted probability of surgery using the main model. X-axis: SFU grade.

Figure 8 - Discrimination Final model - Derivation



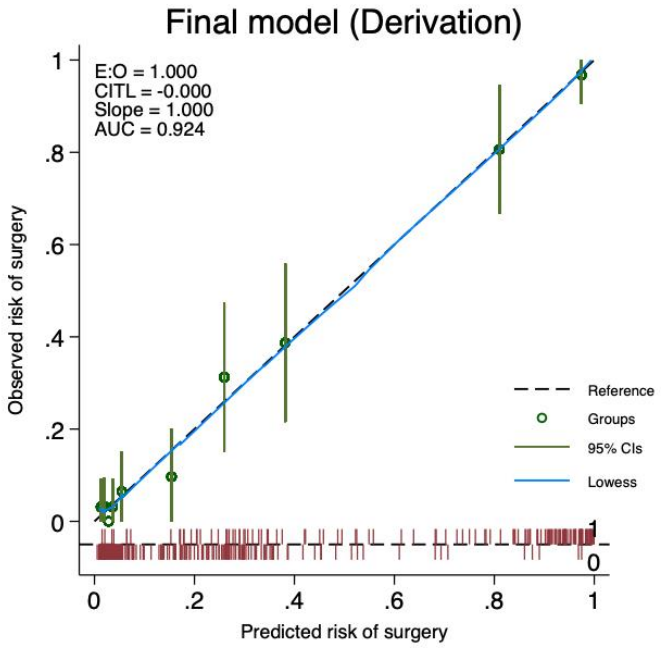
Receiver operating characteristic curve of the derivation cohort indicating the discrimination ability of the final model with natural values in predicting surgery. AUC: Area under the curve. Shows excellent discrimination.

Figure 9- Discrimination final model - Validation



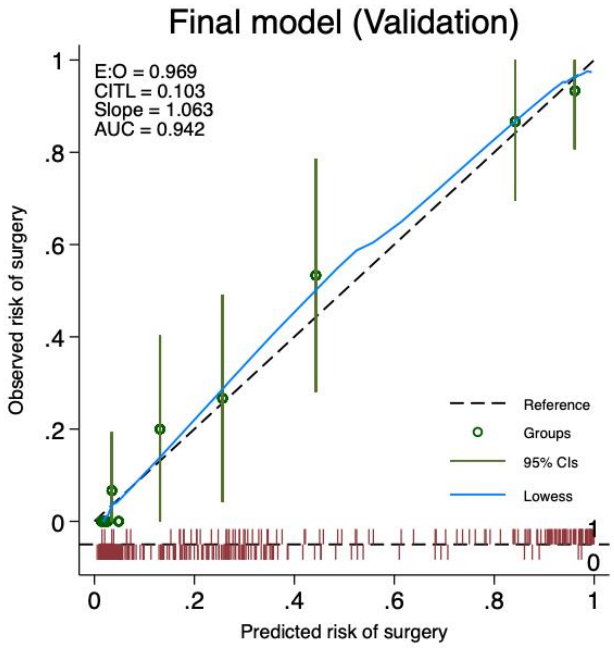
Receiver-operating characteristic curve of the validation cohort indicating the discrimination ability of the final model with natural values in predicting surgery. AUC: Area under the curve. Shows excellent discrimination.

Figure 10 - Calibration plot of main model - Derivation



Calibration plot of the final clinical prediction model using APD, SFU and RLI to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Figure 11- Calibration plot of main model – Validation



Calibration plot of the final clinical prediction model using APD, SFU and RLI as predictors to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Table 8- HERO score cut-off points

	Youden Index	Specificity	Sensitivity
APD cut-offs			
5	0.26	0.29	0.98
10	0.61	0.77	0.84
15	0.57	0.94	0.63
20	0.35	0.99	0.37
RLI cut-offs			
1	0.23	0.31	0.92
5	0.35	0.5	0.84
10	0.47	0.74	0.73
15	0.49	0.89	0.6
SFU cut-offs			
1	0.24	0.25	0.99
2	0.65	0.69	0.95
3	0.57	0.98	0.59
4	0	0.99	0.01

SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter (mm); RLI: Renal length Index (%).

Table 9 - Distribution of the HERO score

SFU grade	**N (%)**	**Distribution**
0	0 (0%)	0 (0%)
1	56 (17.83%)	56 (17.83%)
2	104 (33.12%)	104 (33.12%)
3	99 (31.53%)	99 (31.53%)
4	55 (17.52%)	55 (17.52%)
APD		
< 5mm	43 (13.74%)	3.21±1.12
5 - 10 mm	125 (39.94%)	7.21±1.43
10-15 mm	76 (24.28%)	11.76±1.44
15- 20 mm	29 (9.27%)	17.13±1.45
≥ 20 mm	40 (12.78%)	27.95±8.05
RLI		
100%(affected renal length - contralateral renal length)/affected renal length*		
< 1%	84 (26.75%)	-6.12±7.66
1-5%	46 (14.65%)	3.24±1.08
5-10%	67 (21.34%)	7.14±1.39
10-15%	43 (13.69%)	12.60±1.48
≥ 15%	74 (23.57%)	23.91±7.65
Data are presented as n (%), mean ±standard deviation. Data from derivation cohort (n=314)		
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter (mm); RLI: Renal length index (%).		

Table 10 - Distribution of surgical cases by HERO score categories

SFU grade	**Surgery - No**	**Surgery - Yes**
0		
1	55 (98.21%)	1 (1.79%)
2	100 (96.15%)	4 (3.85%)
3	69 (69.7%)	30 (30.3%)
4	4 (7.27%)	51 (92.73%)
APD		
< 5mm	41 (95.35%)	2 (4.65%)
5 - 10 mm	119 (95.2%)	6 (4.8%)
10-15 mm	51 (67.11%)	25 (32.89%)
15- 20 mm	13 (44.83%)	16 (55.17%)
≥ 20mm	4 (10%)	36 (90%)
RLI		
< 1%	78 (92.86%)	6 (7.14%)
1-5%	39 (84.78%)	7 (15.22%)
5-10%	56 (83.58%)	11 (16.42%)
10-15%	32 (74.42%)	11 (25.58%)
≥ 15%	23 (31.08%)	51 (68.92%)

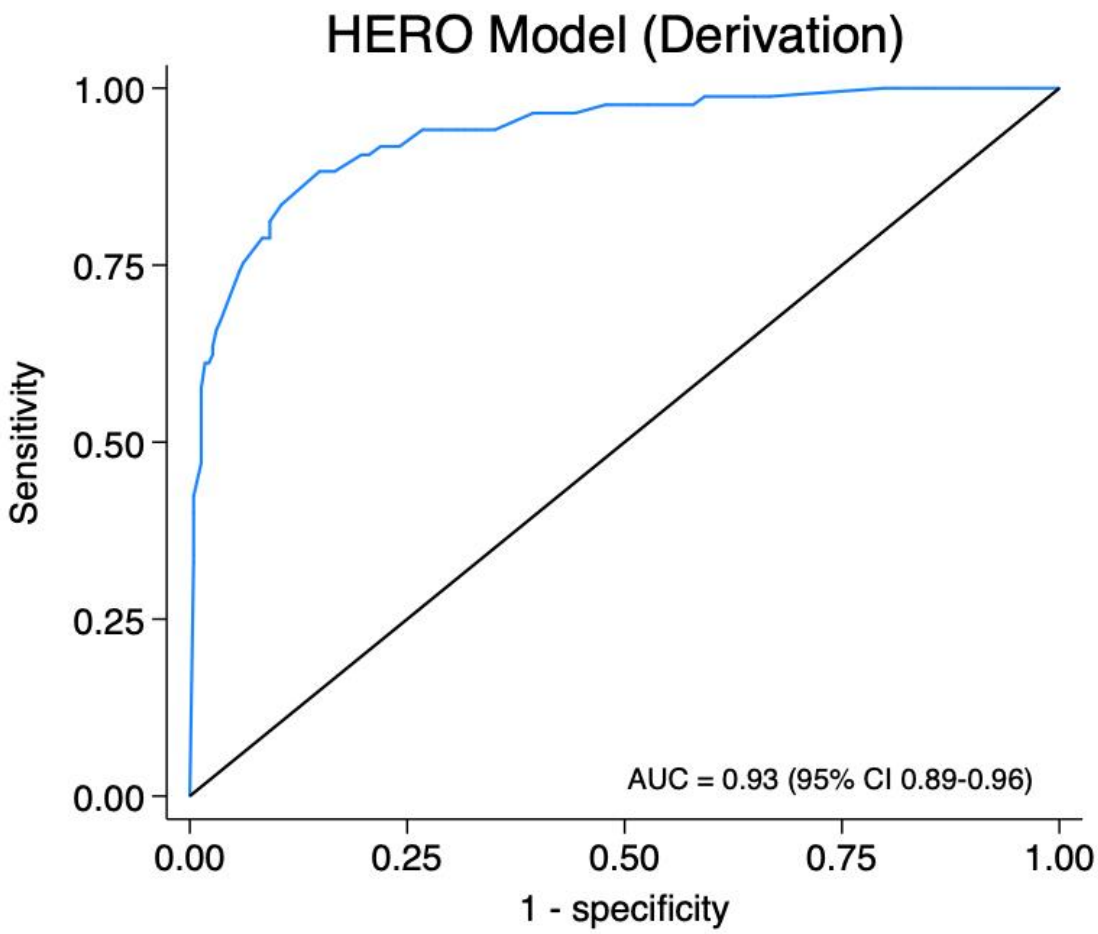
Data are presented as n (%). Data from derivation cohort (n=314).
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter (mm); RLI: Renal length index (%).
Fisher's exact test <0.001 for each category

Table 11 - HERO model to forecast surgery (Derivation)

	Odds ratio	**95% CI**	**p-value**
APD			
5 - 10 mm	Reference		
10.1-15 mm	4.77	1.55-14.68	0.006
15.1- 20 mm	5.33	1.41 - 20.10	0.014
≥ 20 mm	13.29	2.73-64.68	0.001
SFU			
1	Reference		
2	1.84	0.17-19.93	0.614
3	11.55	1.03-129.98	0.048
4	160.96	11.08-2339.0	<0.001
RLI			
1-5%	Reference		
5-10%	0.51	0.14-1.88	0.309
10-15%	0.65	0.18-2.33	0.504
≥15%	1.33	0.37-4.77	0.664

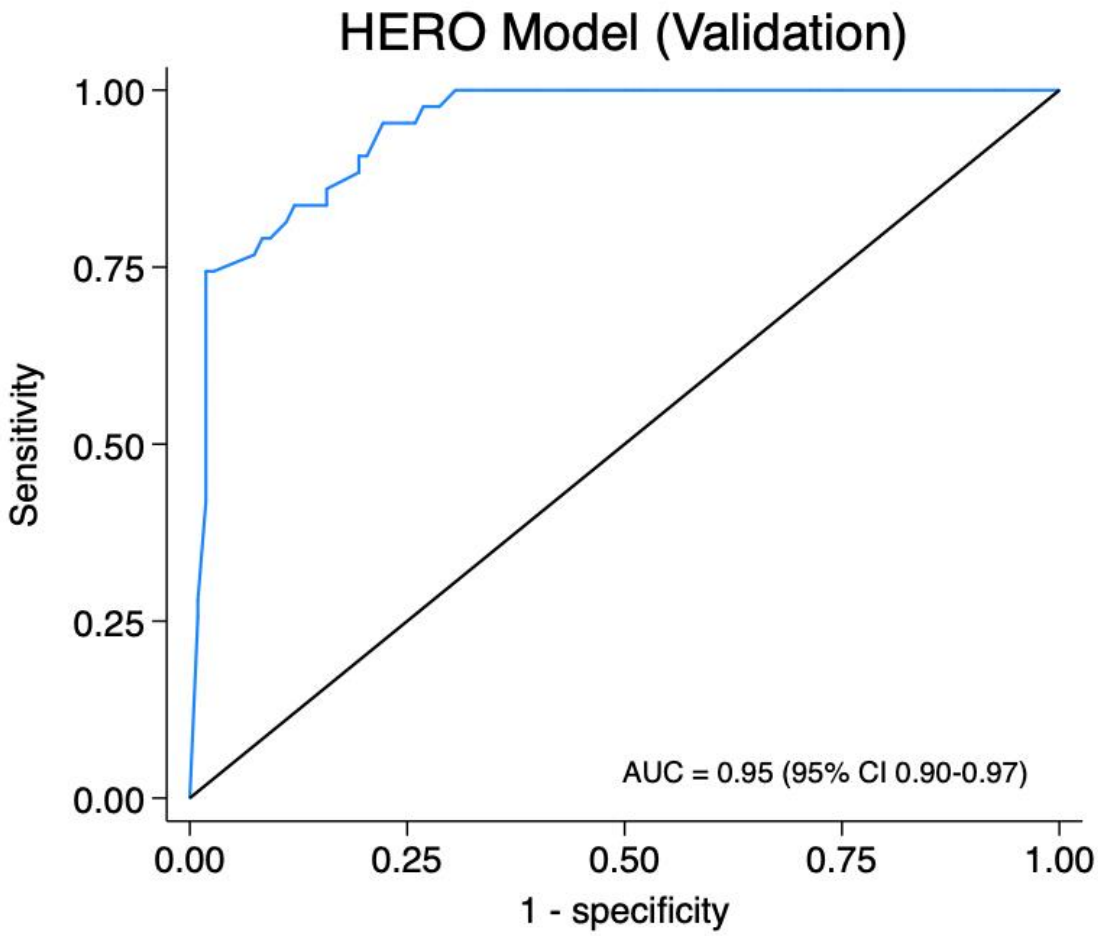
Multivariable logistic regression HERO model. Data from derivation cohort (n=314)
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter (mm); RLI: Renal length index (%); CI: Confidence interval; model AIC: (12df) 197.06

Figure 12 - HERO score discrimination (Derivation)



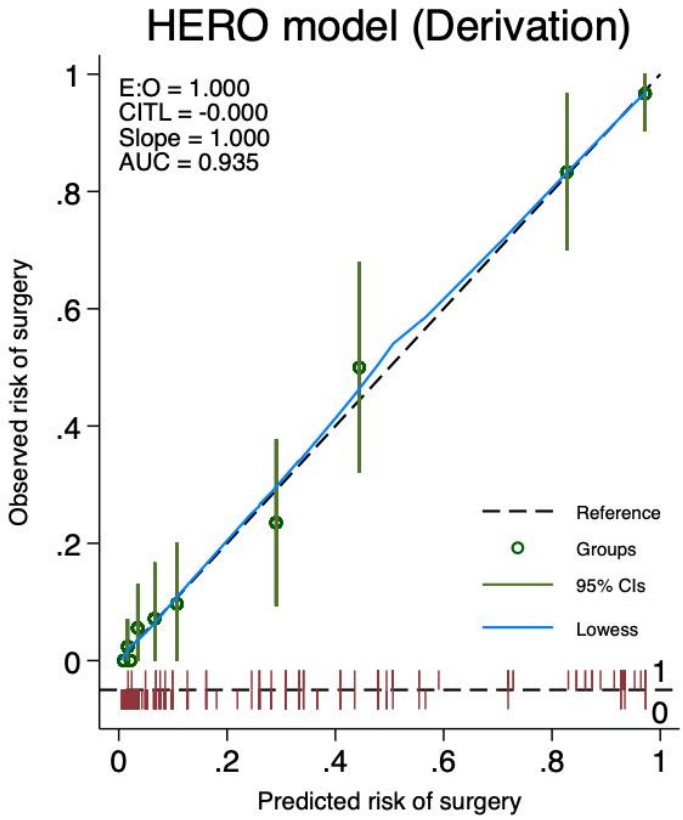
Receiver-operating characteristic curve of the derivation cohort indicating the excellent discrimination ability of the HERO score in predicting surgery. AUC: Area under the curve.

Figure 13 - HERO score discrimination (Validation)



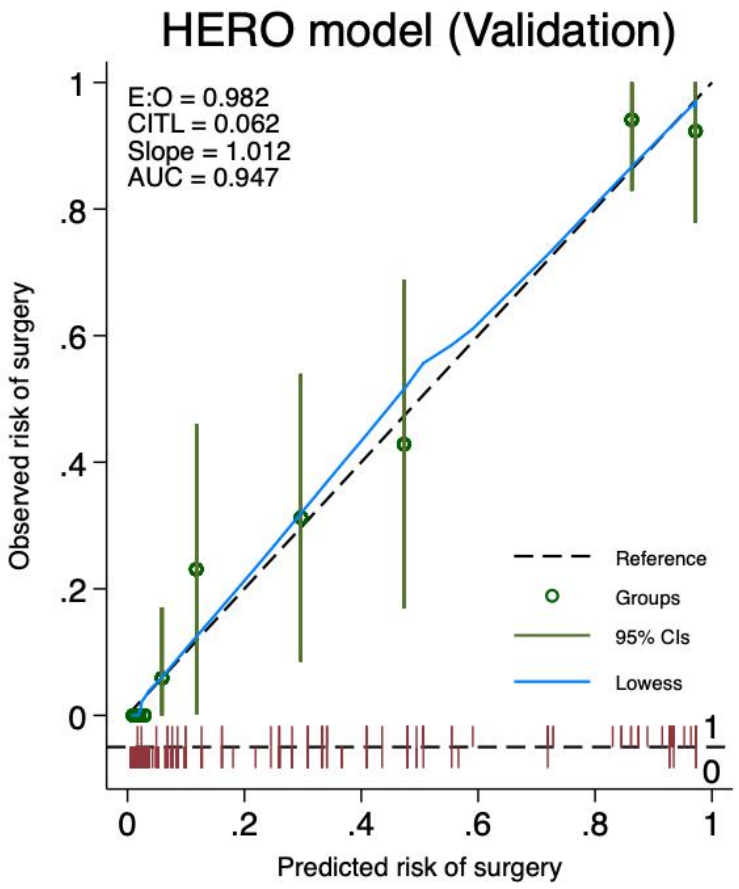
Receiver-operating characteristic curve of the validation cohort indicating the excellent discrimination ability of the HERO score in predicting surgery. AUC: Area under the curve.

Figure 14 - HERO score calibration plot (Derivation)



Calibration plot of the HERO score to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Figure 15- HERO score calibration plot (Validation)



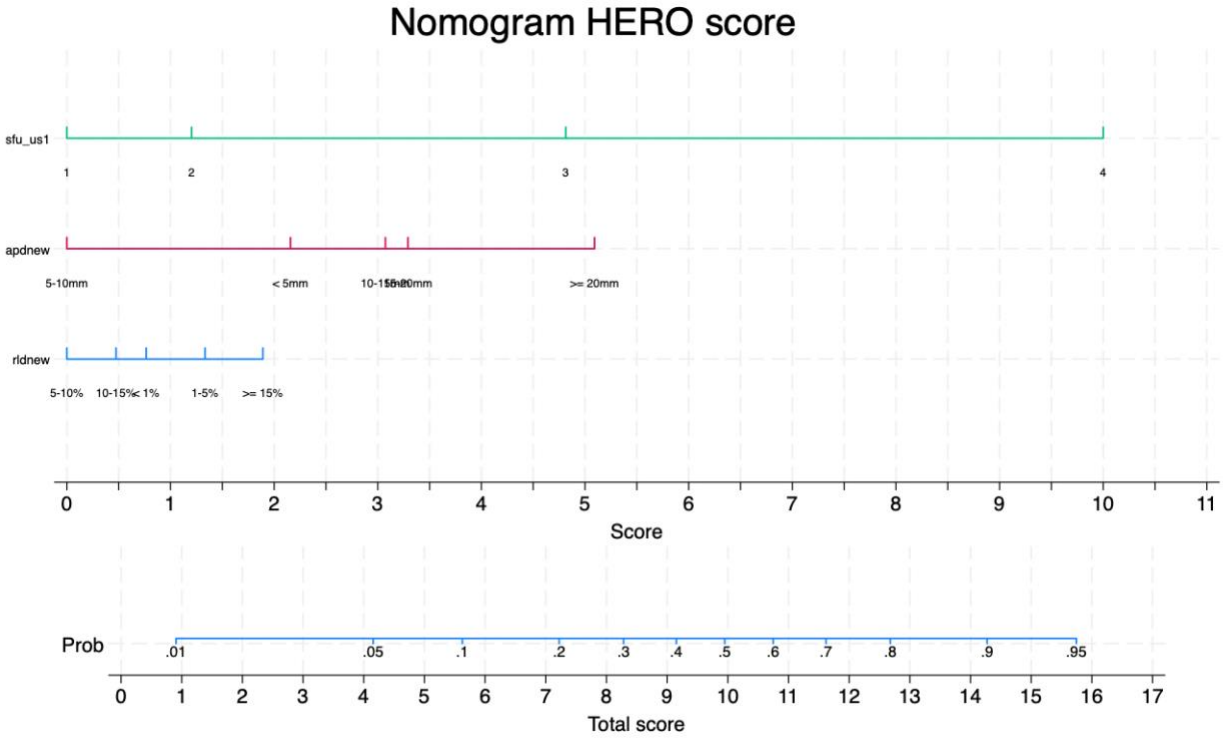
Calibration plot of the HERO score to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Table 12- Simplified HERO model with weighted risk scores

SFU grade	**Log-odds**	**Points**
0	-	0
1	Reference	1
2	0.61	2
3	2.45	5
4	5.08	10
APD		
< 5mm	1.10	0
5 - 10 mm	Reference	1
10-15 mm	1.56	3
15- 20 mm	1.67	4
≥ 20mm	2.59	5
RLI		
< 1%	-0.29	0
1-5%	Reference	0.5
5-10%	-0.68	1
10-15%	-0.44	1.5
≥ 15%	0.28	2

Data from derivation cohort (n=314). Multivariable model LR chi2(11) = 193.03, p<0.0001.
SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter (mm); RLI: Renal length index (%).

Figure 16 - HERO score nomogram



The HERO score nomogram is designed to visualize the simplified HERO scoring system. It features three coloured lines: green for SFU, red for APD, and blue for RLI. Each cut-off point on these lines corresponds to a category.

Scoring Ruler: The first ruler at the bottom displays scores ranging from 0 to 11.

Risk Ruler: The second ruler shows the estimated risk of surgery in a linear fashion.

To use the nomogram: a) Sum the points corresponding to each variable's category. b) Locate this total on the Scoring Ruler. c) Refer to the Risk Ruler to find the estimated surgical risk for the patient.

Table 13 - Surgical risk by simplified weighted HERO score.

Score points	Risk of surgery	Sensitivity	Specificity
1	0.8%	1.00	0.05
1.5	1.0%	1.00	0.07
2	1.3%	0.99	0.18
2.5	1.6%	0.98	0.21
3	2.1%	0.96	0.40
3.5	2.7%	0.96	0.46
4	3.4%	0.96	0.54
4.5	4.3%	0.96	0.57
5	5.5%	0.96	0.61
5.5	6.9%	0.94	0.63
6	8.7%	0.94	0.69
6.5	10.9%	0.94	0.72
7	13.5%	0.92	0.77
7.5	16.7%	0.92	0.79
8	20.5%	0.88	0.80
8.5	24.8%	0.85	0.82
9	29.7%	0.82	0.86
9.5	35.2%	0.76	0.91
10	41.1%	0.69	0.95
10.5	47.2%	0.67	0.96
11	53.4%	0.64	0.97
11.5	59.5%	0.62	0.98
12	65.4%	0.58	0.98
12.5	70.8%	0.58	0.99
13	75.6%	0.56	0.99
14	83.6%	0.54	0.99
14.5	86.8%	0.52	0.99
15	89.4%	0.44	1.00
15.5	91.5%	0.41	1.00
16	93.3%	0.33	1.00

17	95.8%	0.00	1.00

Table 14 - Distribution of surgical cases by simplified weighted HERO score

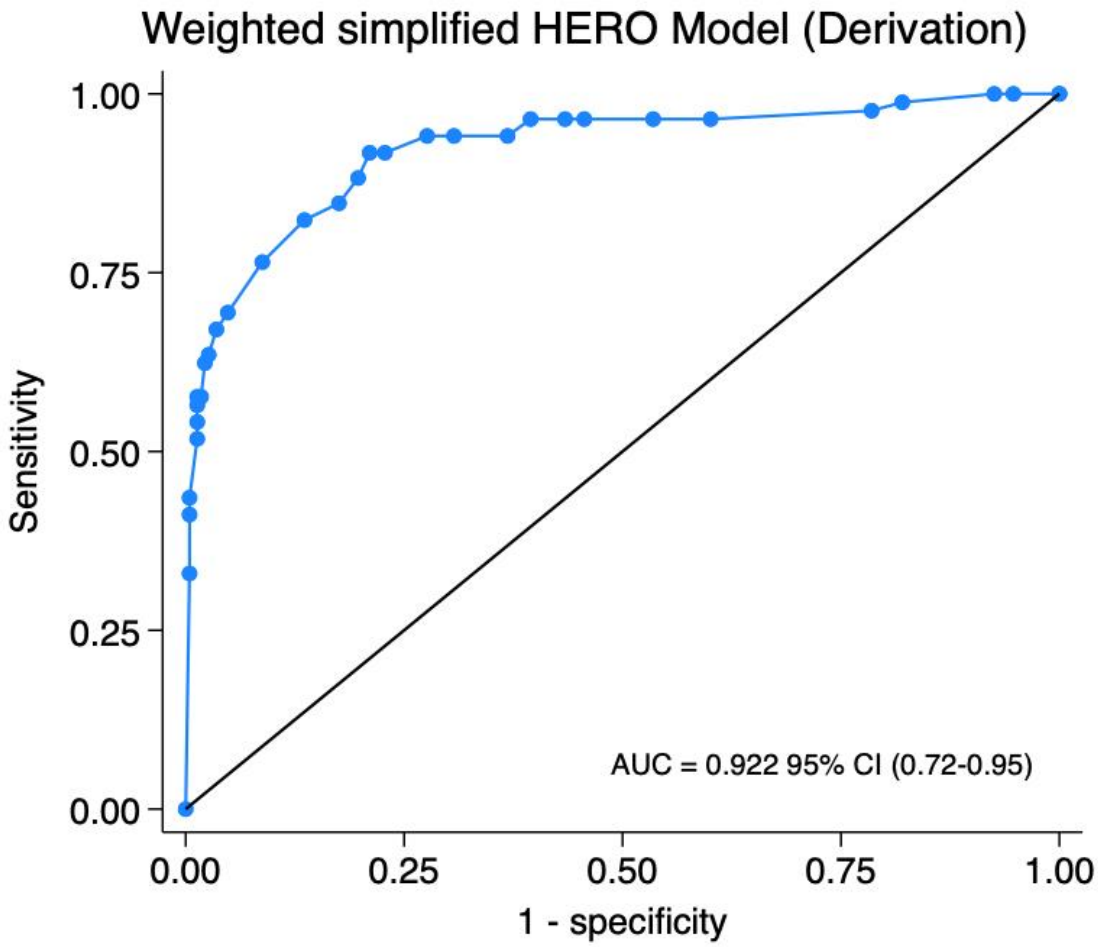
Variable	Surgery - No	Surgery - Yes
1	12 (100%)	0 (0%)
1.5	5 (100%)	0 (0%)
2	24 (96%)	1 (4%)
2.5	8 (88.89%)	1 (11.11%)
3	42 (97.67%)	1 (2.33%)
3.5	15 (100%)	0 (0%)
4	18 (100%)	0 (0%)
4.5	5 (100%)	0 (0%)
5	9 (100%)	0 (0%)
5.5	6 (75%)	2 (25%)
6	14 (100%)	0 (0%)
6.5	7 (100%)	0 (0%)
7	11 (84.62%)	2 (15.38%)
7.5	4 (100%)	0 (0%)
8	3 (50%)	3 (50%)
8.5	5 (62.5%)	3 (37.5%)
9	9 (81.82%)	2 (18.18%)
9.5	11 (68.75%)	5 (31.25%)
10	9 (60%)	6 (40%)
10.5	3 (60%)	2 (40%)
11	2 (40%)	3 (60%)
11.5	1 (50%)	1 (50%)
12	1 (20%)	4 (80%)
12.5	1 (100%)	0 (0%)
13	0 (0%)	1 (100%)
14	0 (0%)	2 (100%)
14.5	0 (0%)	2 (100%)
15	2 (22.22%)	7 (77.78%)
15.5	0 (0%)	2 (100%)
16	0 (0%)	7 (100%)
17	1 (3.45%)	28 (96.55%)

Data are presented as n (%). Data from derivation cohort (n=314). Pearson chi2(30) = 190.221, p<0.001; Fisher's exact test not supported

Table 15 - Grouped Surgical risk by simplified weighted HERO score

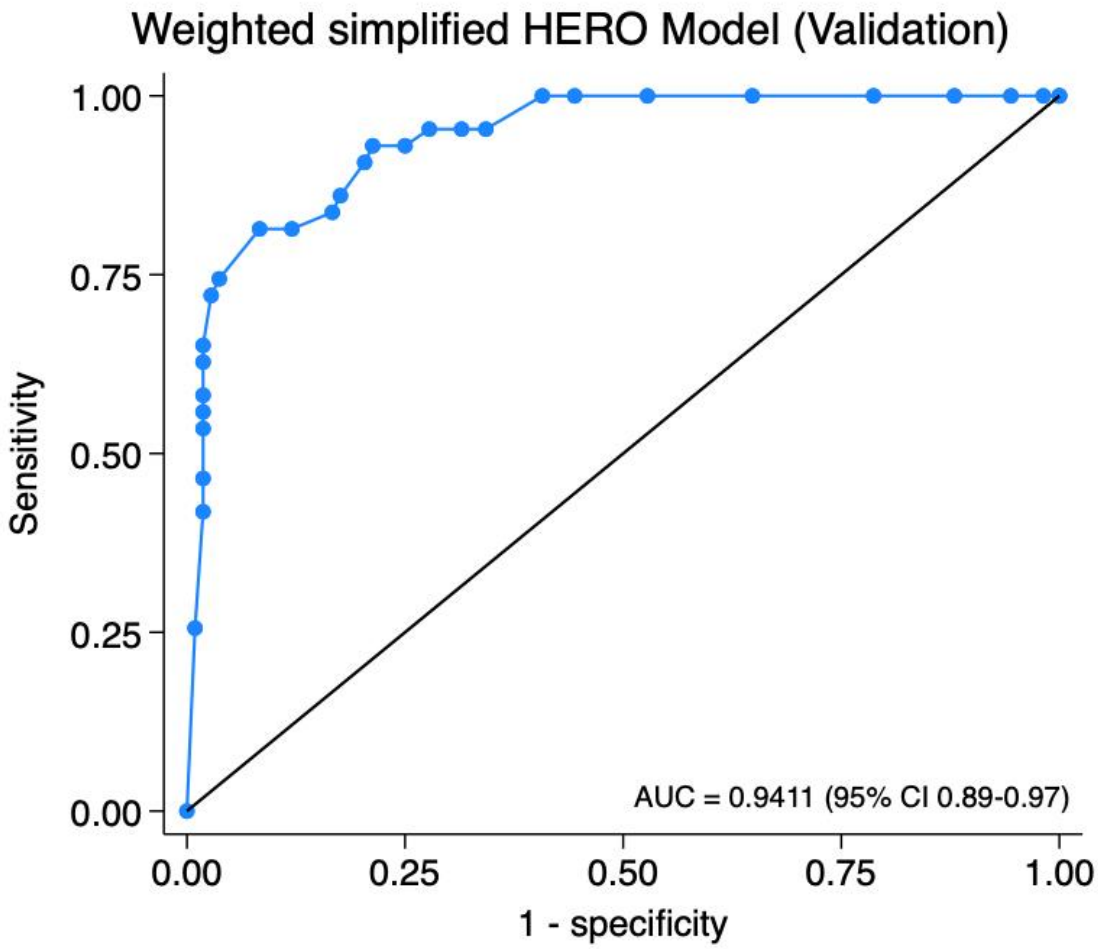
Score points	Risk of surgery
0-6	8.7%
6.5-8.5	17.3%
9-10.5	38.3%
11-12.5	62.3%
13-14.5	82.0%
15-17	92.5%

Figure 17- simplified HERO score discrimination plot (Derivation)



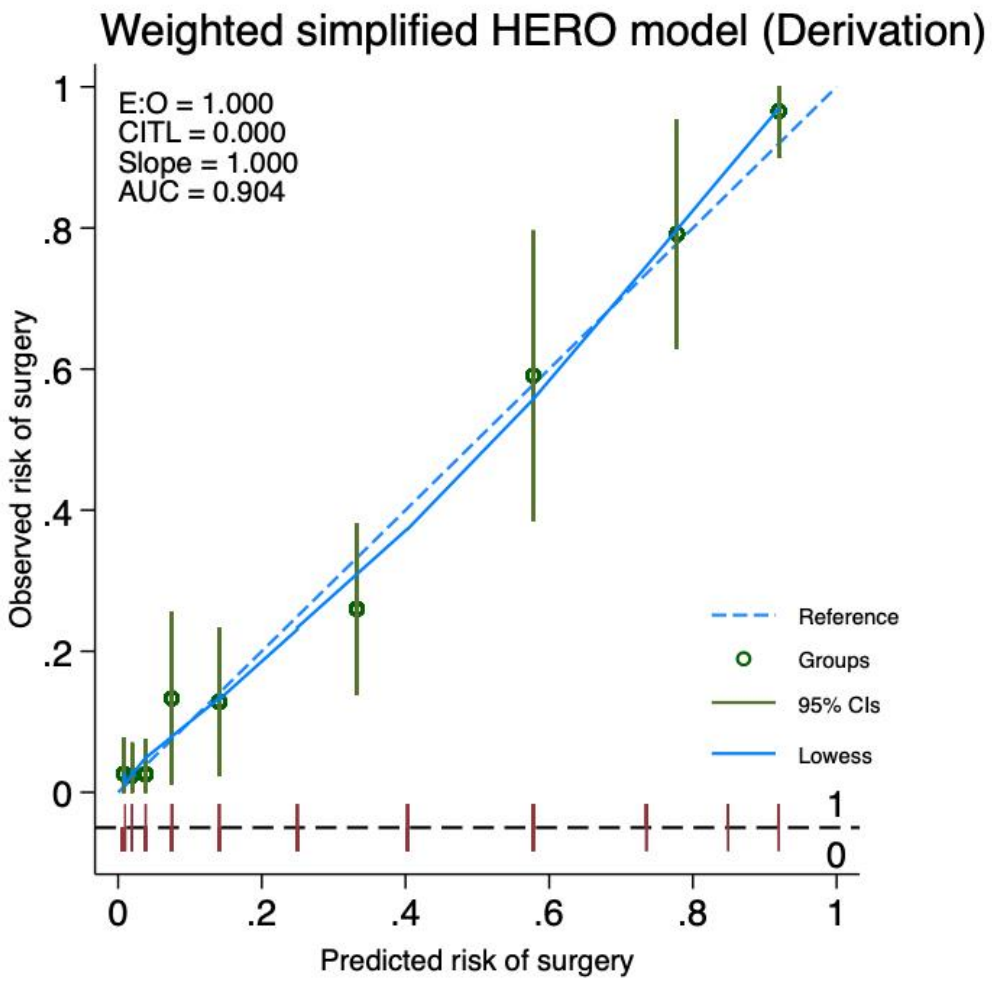
Discrimination plot: Receiver-operating characteristic curve of the derivation cohort indicating the excellent discrimination ability of the simplified weighted HERO score in predicting surgery. AUC: Area under the curve.

Figure 18 - simplified HERO score discrimination plot (Validation)



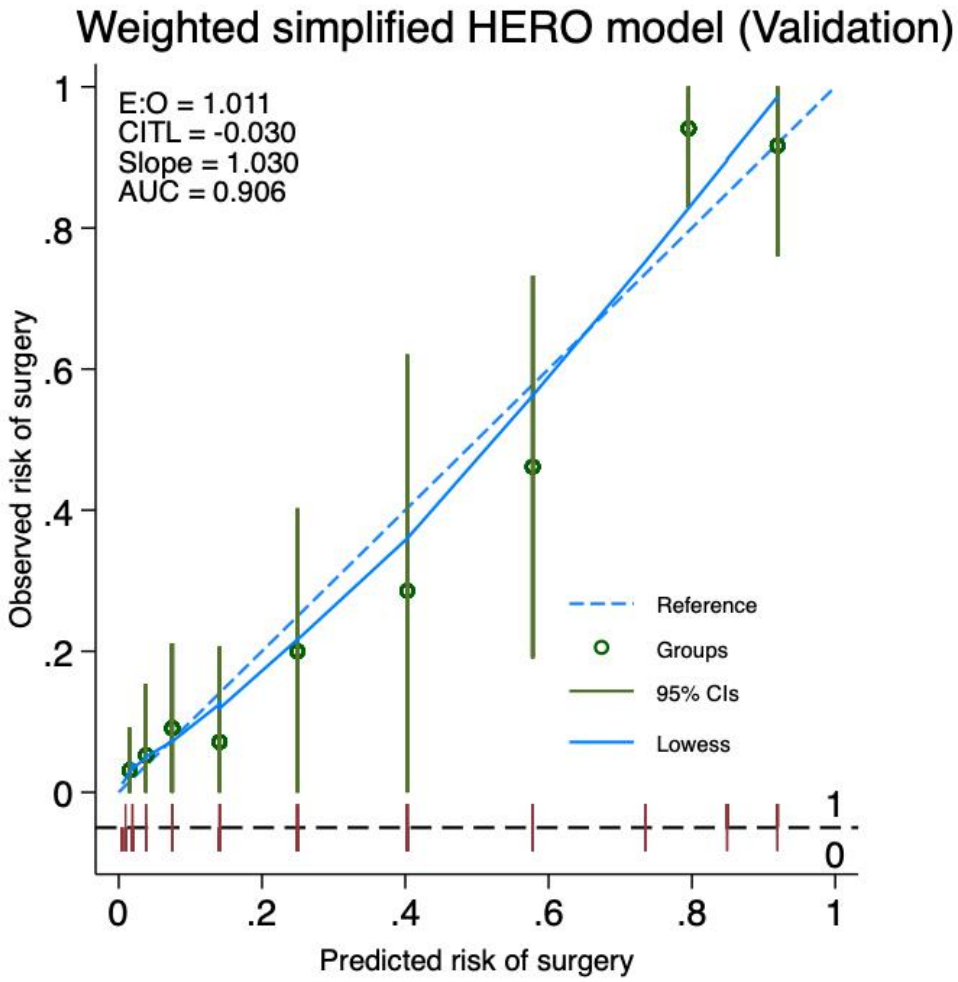
Discrimination plot: Discrimination plot: Receiver-operating characteristic curve of the derivation cohort indicating the excellent discrimination ability of the simplified weighted HERO score in predicting surgery. AUC: Area under the curve.

Figure 19- simplified HERO score calibration plot (Derivation)



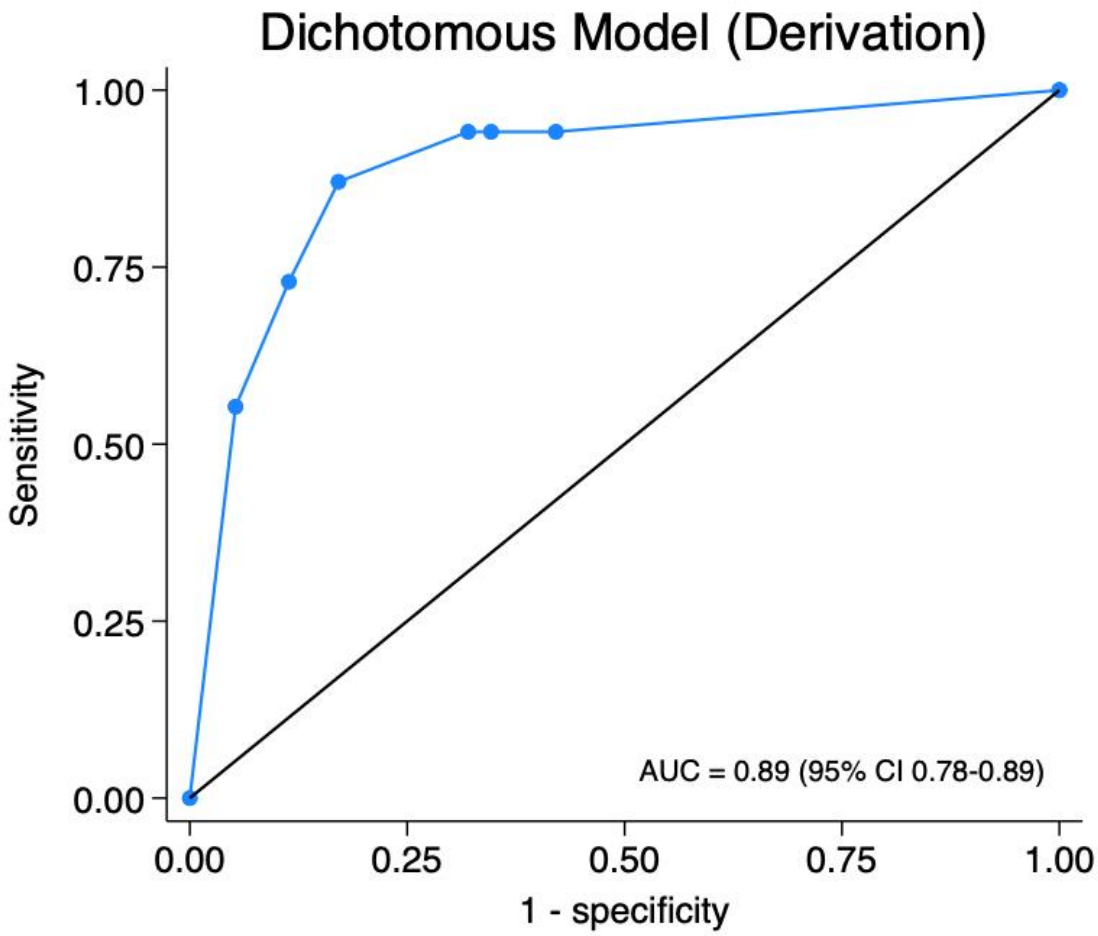
Calibration plot of the simplified weighted HERO applied to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Figure 20- simplified HERO score calibration plot (Validation)



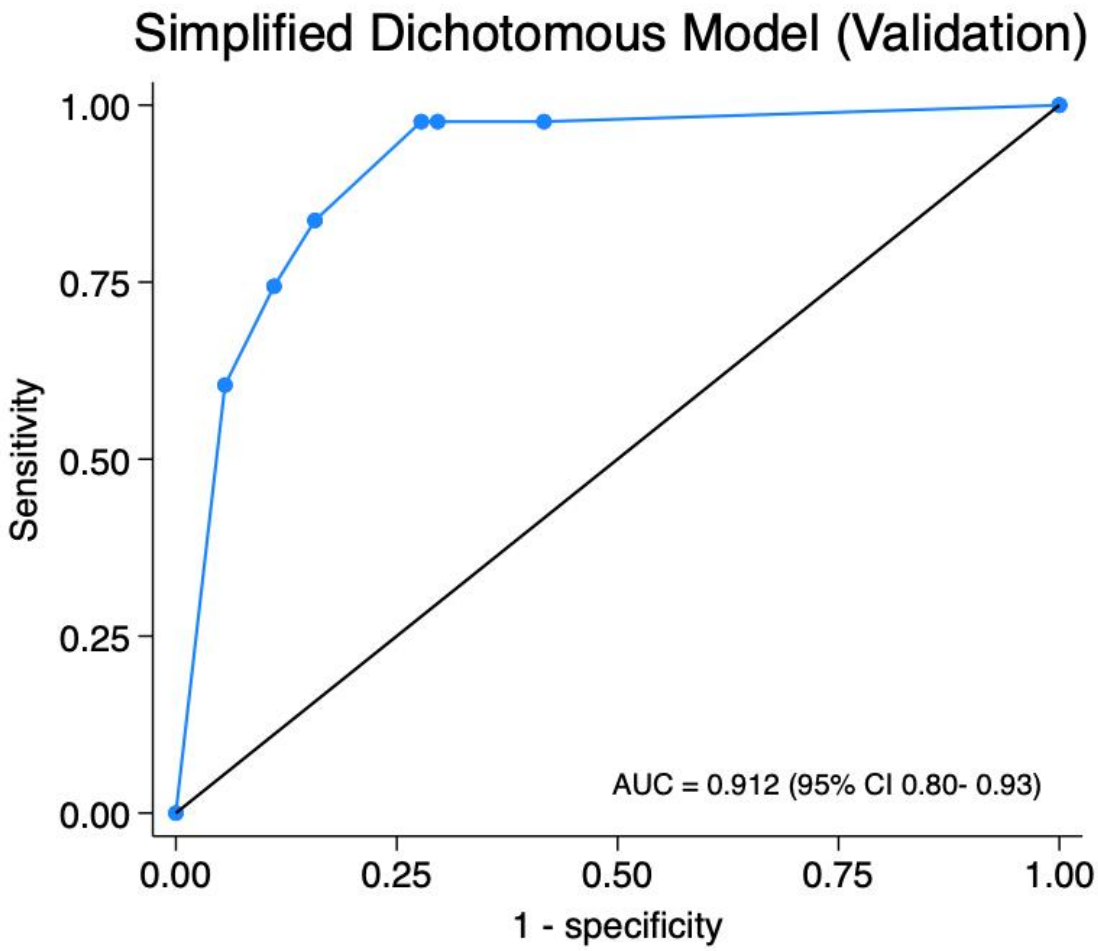
Calibration plot of the simplified weighted HERO applied to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Figure 21- Discrimination plot dichotomous model (Derivation)



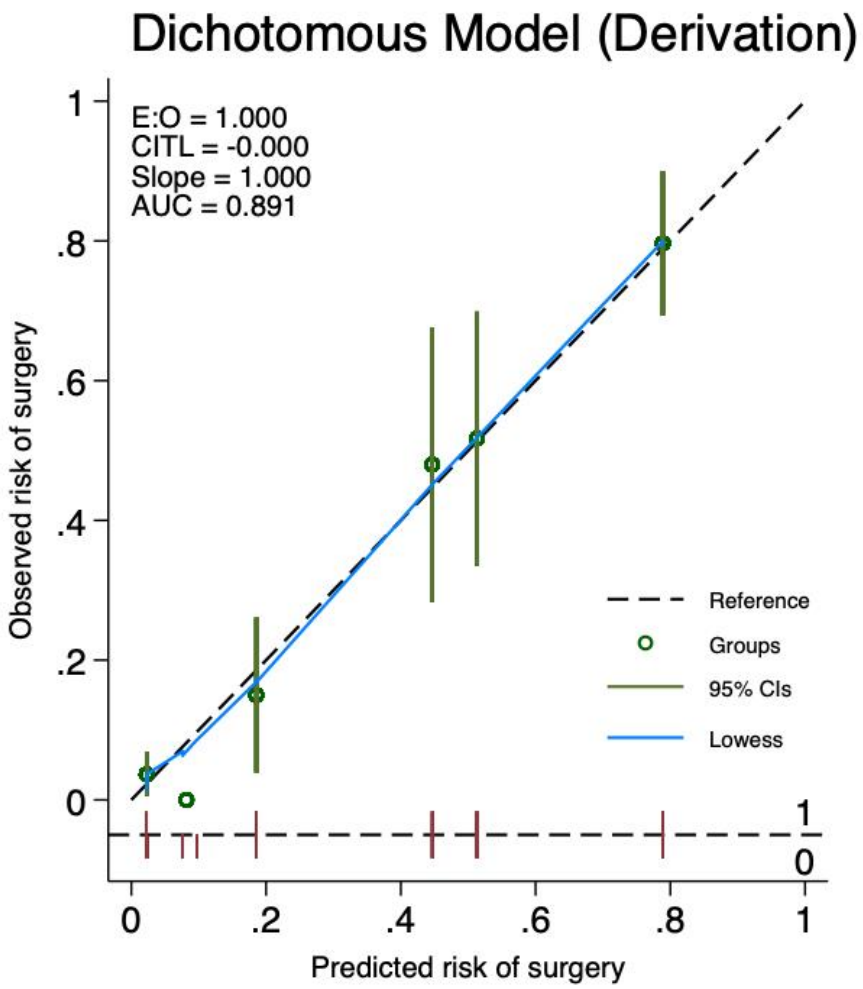
Discrimination plot: Receiver-operating characteristic curve of the derivation cohort indicating the good discrimination ability of the dichotomized score in predicting surgery. AUC: Area under the curve

Figure 22- Discrimination plot dichotomous model (Validation)



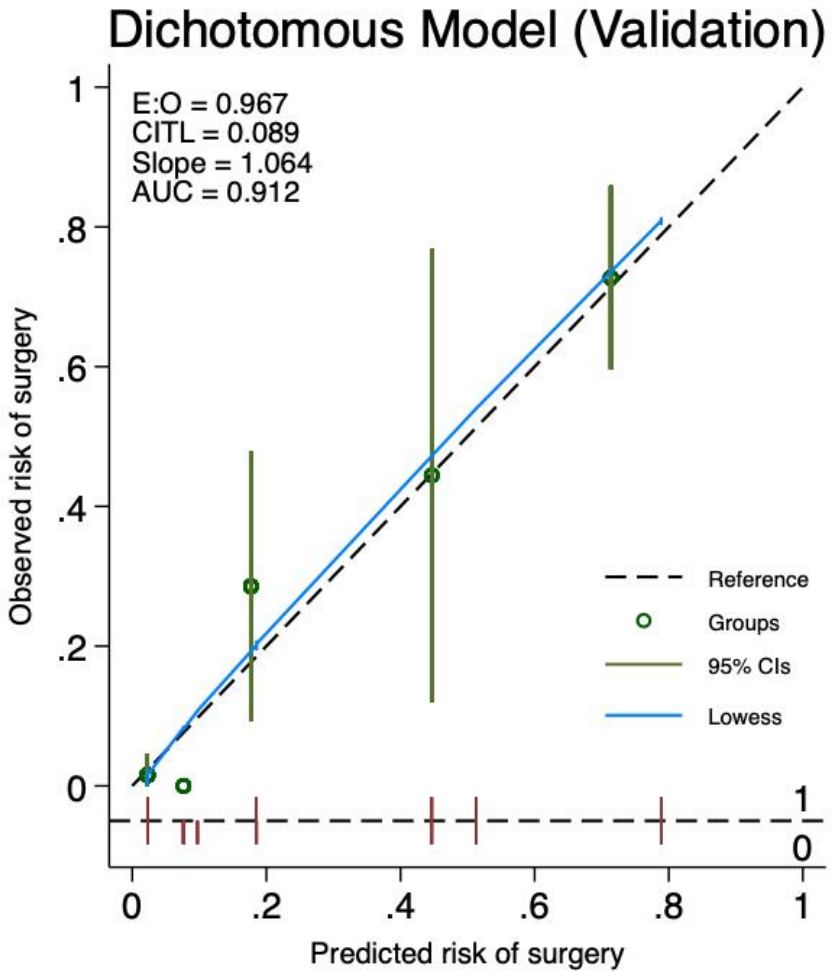
Discrimination plot: Receiver-operating characteristic curve of the validation cohort indicating the good discrimination ability of the dichotomized score in predicting surgery. AUC: Area under the curve

Figure 23- Calibration plot dichotomous model (Derivation)



Calibration plot of the dichotomous model applied to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Figure 24 - Calibration plot dichotomous model (Validation)



Calibration plot of the simplified dichotomous model applied to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Table 16 - HERO score model using dichotomous variables (n=313)

Independent variable	Adjusted OR	95% CI	p-value
SFU			
1-2	Reference		
3-4	9.77	3.45-27.68	<0.001
APD, mm			
< 13	Reference		
≥ 13	4.63	2.27-9.43	<0.001
RLI, %			
<12	Reference		
≥ 12	3.54	1.78-7.05	<0.001

Sensitivity analysis: Multivariable logistic regression model using dichotomous variables for the HERO score. SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length Index. AIC: (4df) 224.52

Table 17- Dichotomized simplified model with risk scores

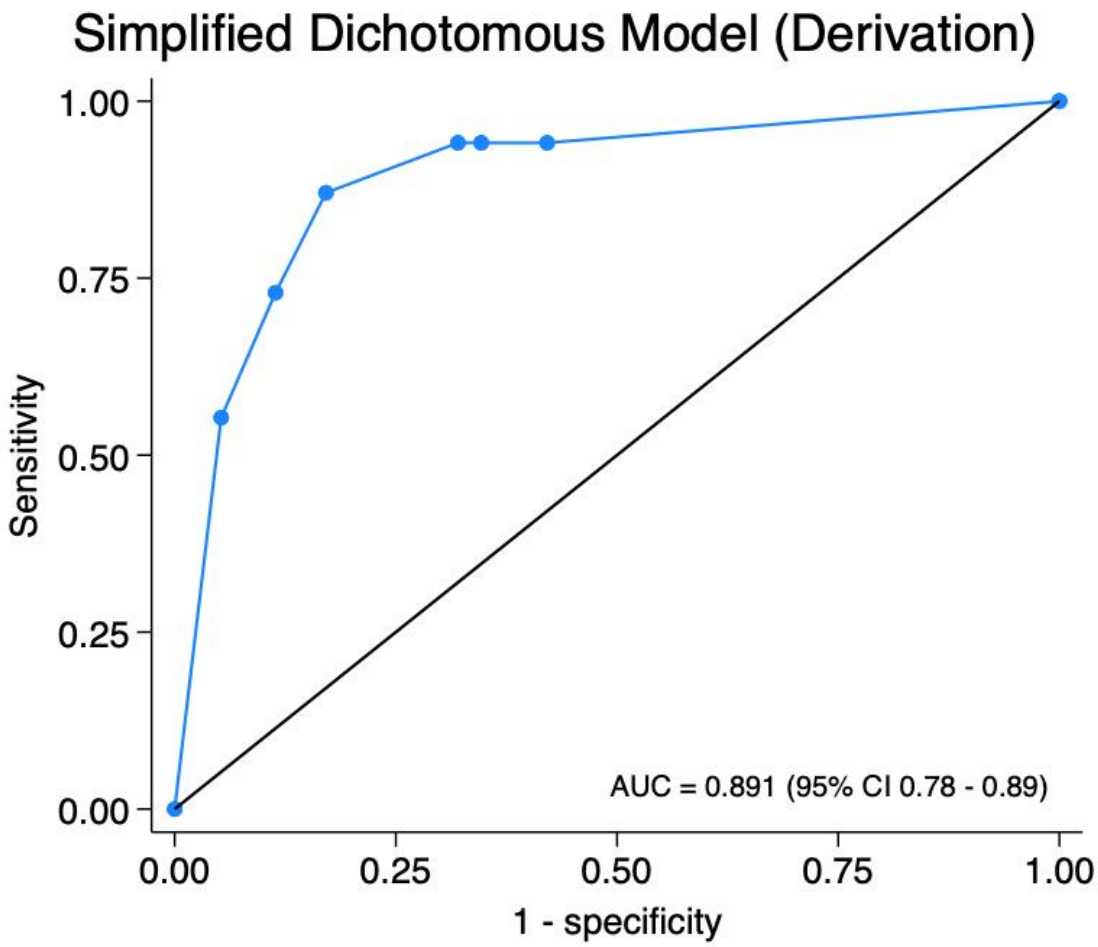
Variable	Categories	Odds Ratio	Points
SFU 3-4	Yes	9.77	10
	No	1	0
APD ≥ 13 mm	Yes	4.63	5
	No	1	0
RLI ≥12%	Yes	3.54	4
	No	1	0

SFU: Society for Fetal Urology grading system; APD: Anteroposterior pelvic diameter; RLI: Renal length Index.

Table 18 - Simplified dichotomized score and predicted probability of surgery

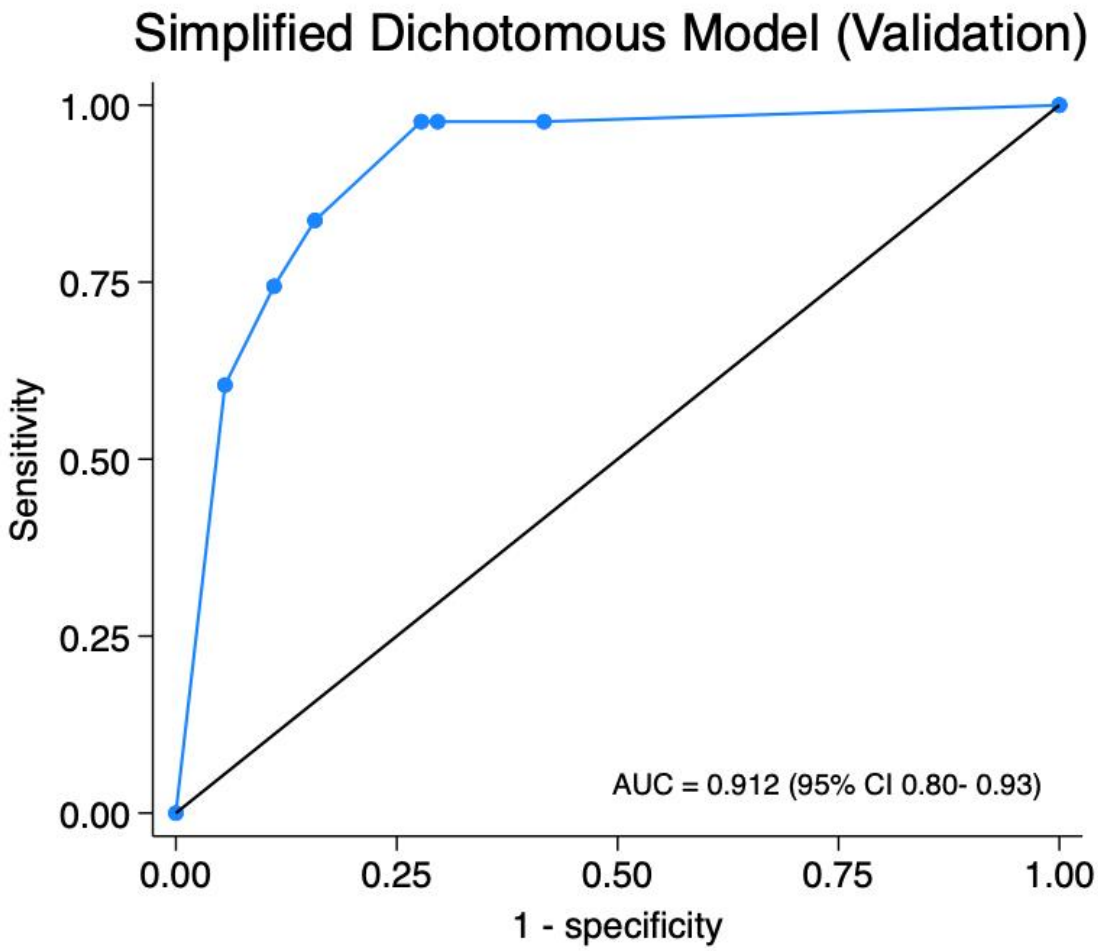
Score points	Risk of surgery	Sensitivity	Specificity
0	1.9%	0.94	0.58
4	5.4%	0.94	0.65
5	7.0%	0.94	0.68
10	22.5%	0.87	0.83
14	46.0%	0.73	0.89
15	52.8%	0.55	0.95
19	76.7%	0.00	1.00

Figure 25 - Discrimination plot simplified dichotomous model (Derivation)



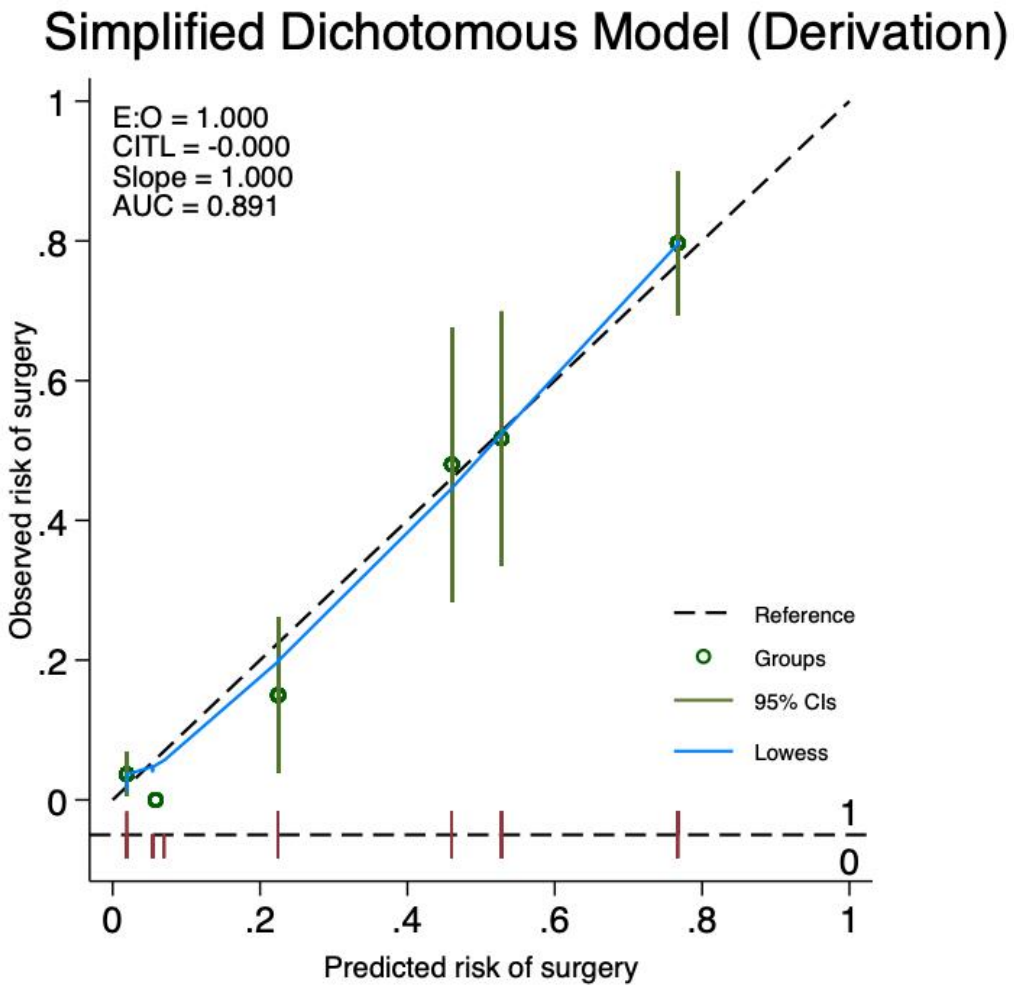
Receiver-operating characteristic curve of the derivation cohort indicating great discrimination ability of the simplified dichotomous model in predicting surgery. AUC: Area under the curve.

Figure 26- Discrimination plot simplified dichotomous model (Validation)



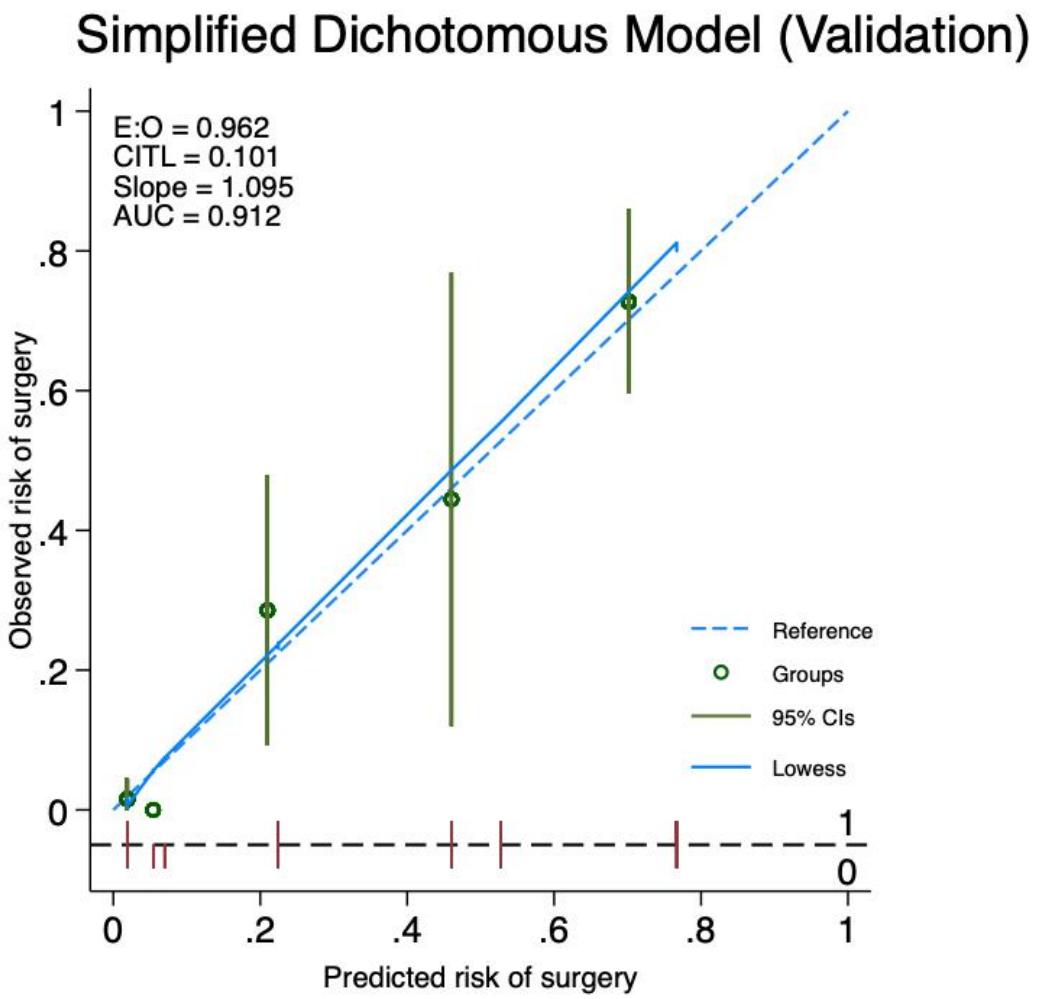
Receiver-operating characteristic curve of the validation cohort indicating excellent discrimination ability of the simplified dichotomous model in predicting surgery. AUC: Area under the curve.

Figure 27 - Calibration plot simplified dichotomous model (Derivation)



Calibration plot of the simplified dichotomous score to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes. The Hosmer-Lemeshow test showed good fit (chi2(4) = 5.26, p = 0.262).

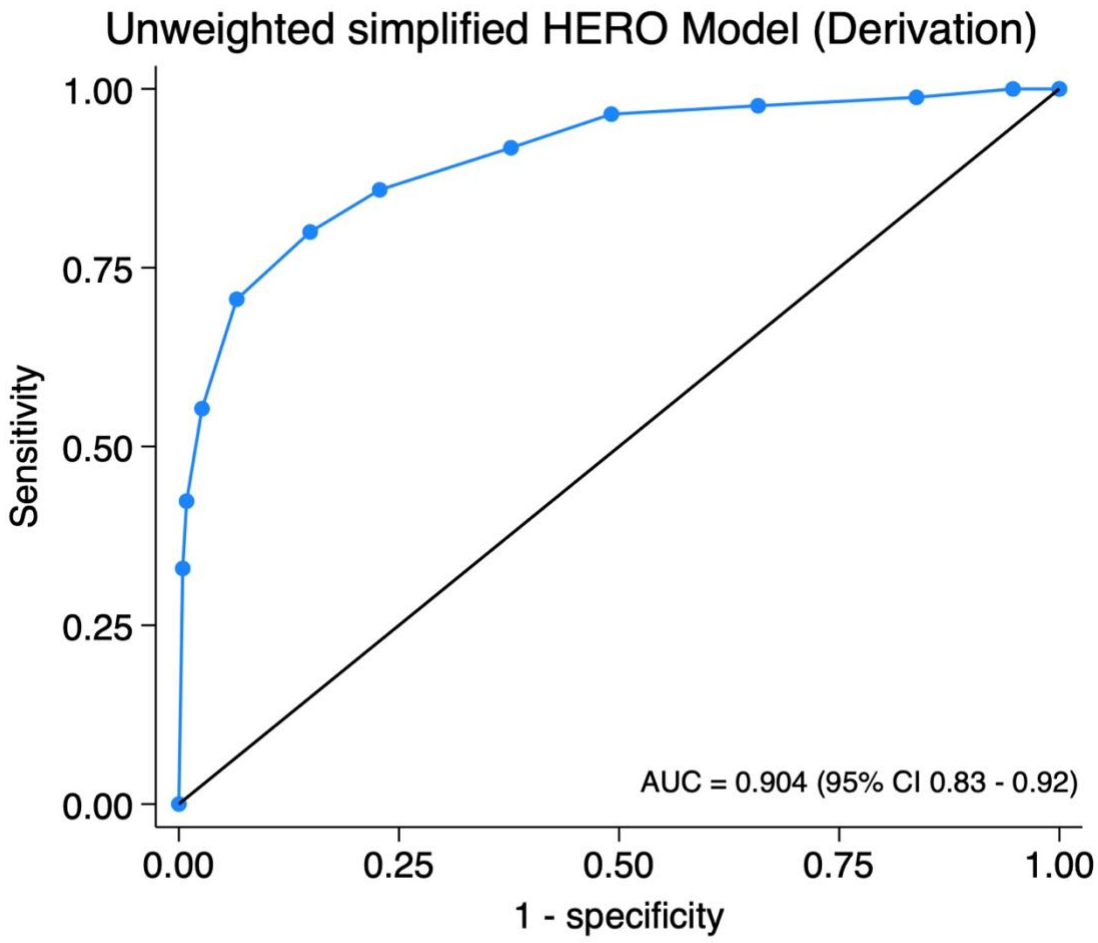
Figure 28 - Calibration plot simplified dichotomous model (Validation)



Calibration plot of the simplified dichotomous score applied to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes. The Hosmer-Lemeshow test showed a good fit (chi2(3) = 1.42, p = 0.691).

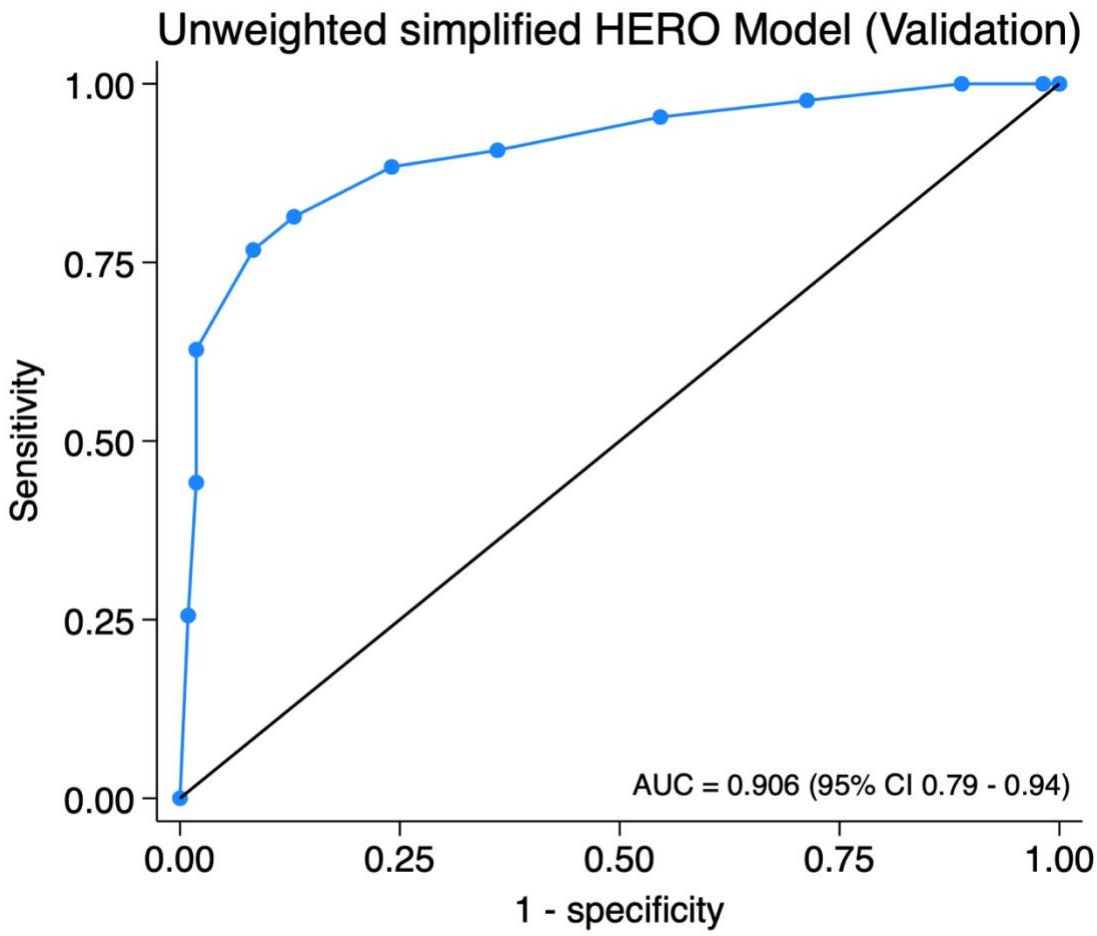
SENSITIVITY ANALYSIS

Figure 29 - Discrimination plot simplified unweighted HERO model (Derivation)



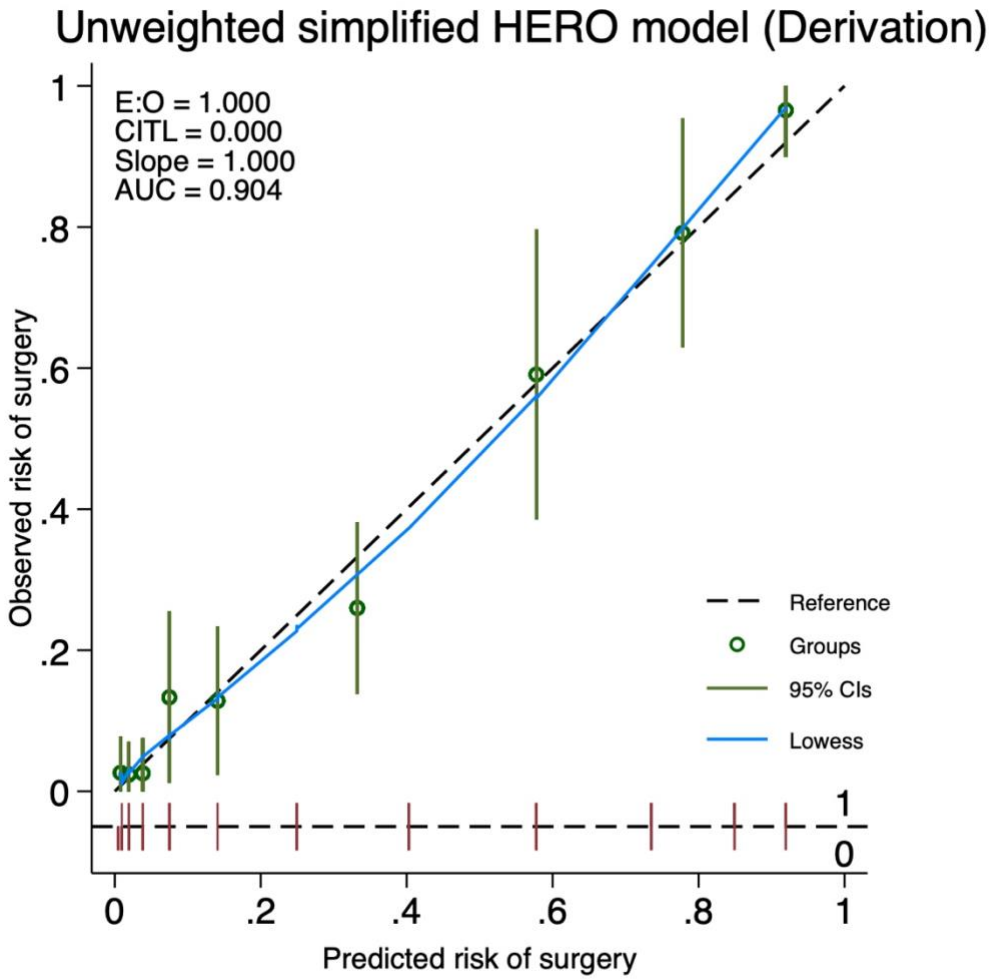
Receiver-operating characteristic curve of the derivation cohort indicating excellent discrimination ability of the simplified unweighted HERO model in predicting surgery. AUC: Area under the curve

Figure 30- Discrimination plot simplified unweighted HERO model (Validation)



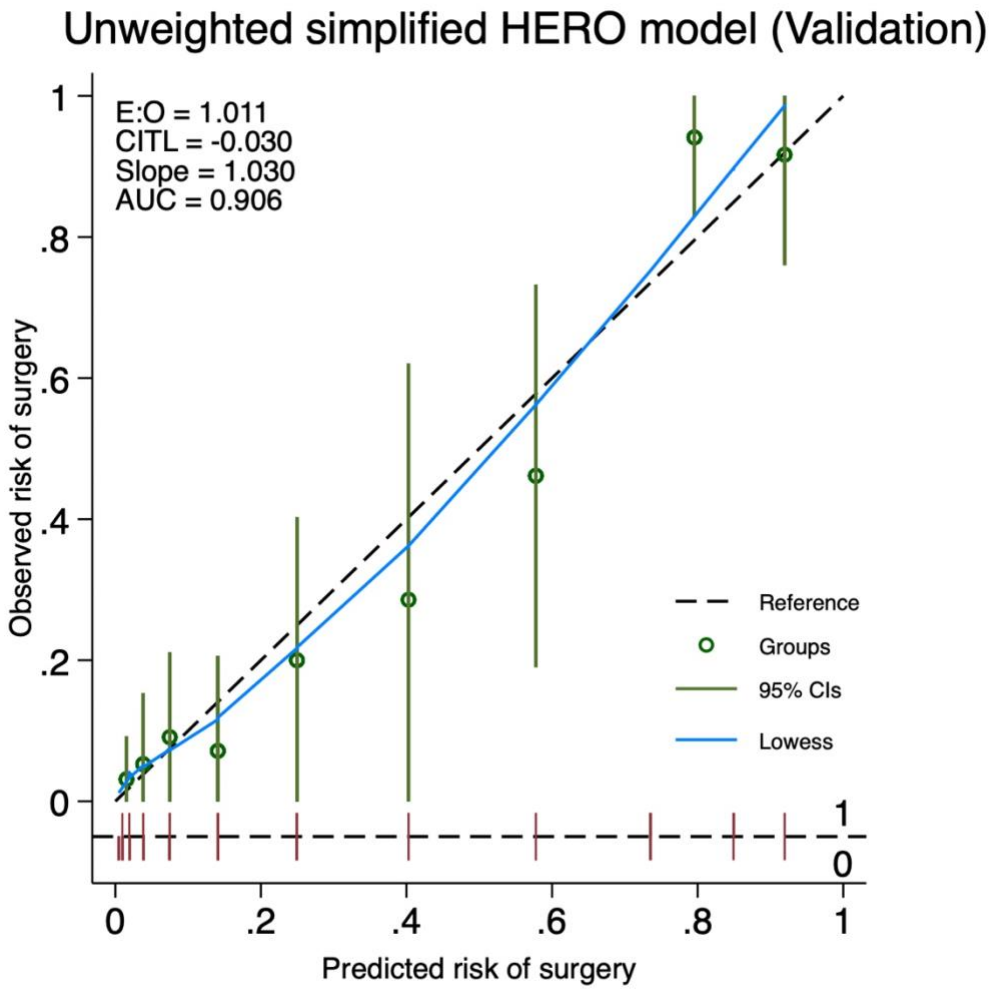
Receiver-operating characteristic curve of the derivation cohort indicating excellent discrimination ability of the simplified unweighted HERO model in predicting surgery. AUC: Area under the curve

Figure 31 - Calibration plot simplified unweighted HERO model (Derivation)



Calibration plot of the simplified unweighted HERO applied to the derivation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes. The Hosmer-Lemeshow test showed good fit (chi2(9) = 5.39, p= 0.612

Figure 32 -Calibration plot simplified unweighted HERO model (Validation)



Calibration plot of the simplified unweighted HERO score applied to the validation cohort. Y-axis: observed proportion of patients undergoing surgery. X-axis: predicted risk of surgery. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

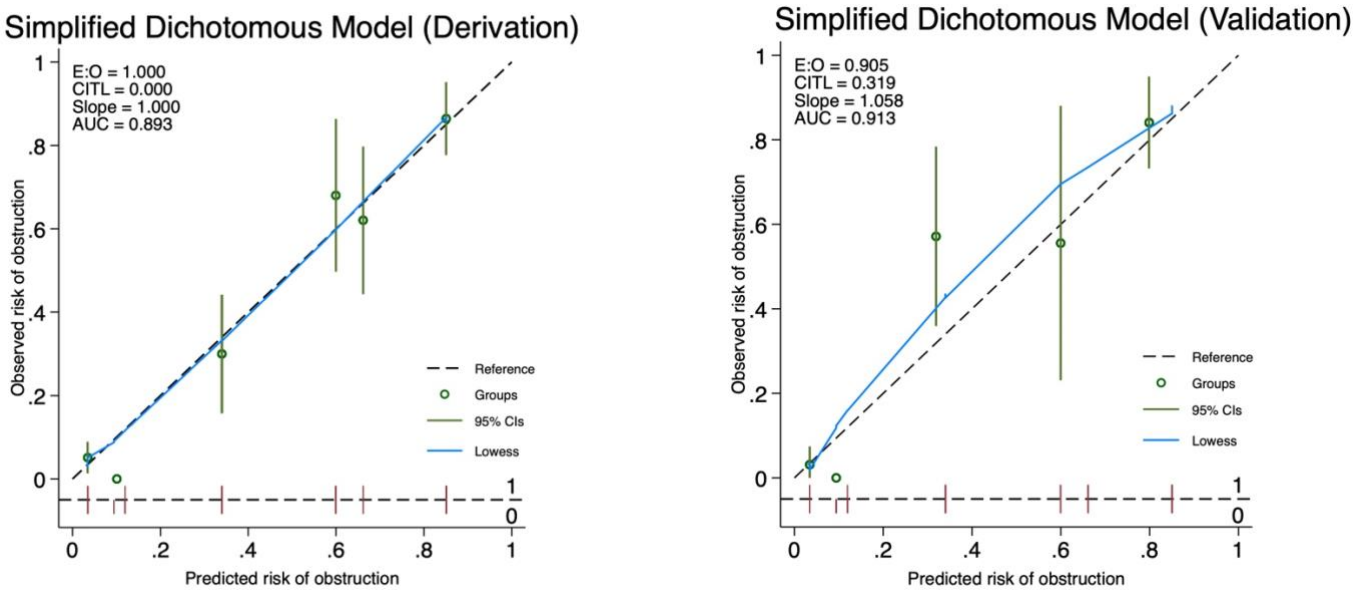
Table 19 - Surgical risk by simplified unweighted HERO score

Unweighted score	Frequency	Risk of surgery (%)
1	12 (3.83%)	0.5%
2	26 (8.31%)	1.0%
3	42 (13.42%)	1.9%
4	39 (12.46%)	3.8%
5	30 (9.58%)	7.5%
6	39 (12.46%)	14.1%
7	23 (7.35%)	24.9%
8	27 (8.63%)	40.3%
9	22 (7.03%)	57.8%
10	15 (4.79%)	73.5%
11	9 (2.88%)	84.9%
12	29 (9.27%)	92.0%

Table 20- Distribution of risk of renal obstruction by simplified dichotomous score

Simplified dichotomous score	Non-obstructed	Obstructed	Risk of obstruction
0	130 (94.89%)	7 (5.11%)	3.45%
4	17 (100%)	0 (0%)	9.4%
5	6 (100%)	0 (0%)	11.9%
10	28 (70%)	12 (30%)	34.0%
14	8 (32%)	17 (68%)	60.0%
15	11 (37.93%)	18 (62.07%)	66.2%
19	8 (13.56%)	51 (86.44%)	85.0%

Figure 33- Calibration plot of simplified dichotomized model and obstruction



A) B)

Calibration plot of the simplified dichotomous model applied to the A) Derivation and B) validation cohort. Y-axis: observed proportion of patients with obstruction on the renal scan. X-axis: predicted risk of renal scan obstruction. Reference: line representing perfect prediction risk equal to observed proportion of patients. Lowess: non-parametric curve representing the plotted observed vs predicted outcomes.

Table 21 - TRIPOD checklist for Prediction Model Development and Validation

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	i
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	iv
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1-11
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	12
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	13
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	13-15
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	13-14
	5b	D;V	Describe eligibility criteria for participants.	13
	5c	D;V	Give details of treatments received, if relevant.	13-15
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	17
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	N/A
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	18
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	N/A
Sample size	8	D;V	Explain how the study size was arrived at.	18
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	23
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	17-21
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	17-21
	10c	V	For validation, describe how the predictions were calculated.	20-21
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	17-21
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	N/A
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	20-21
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	13-21
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	25; 59
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	25;59
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	25;59
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	27-35
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	27-35
	15b	D	Explain how to the use the prediction model.	40-44
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	27-35
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	N/A
Discussion				

Limitations	18	);V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	47-48
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	42-46
	19b	);V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	42-46
Implications	20	);V	Discuss the potential clinical use of the model and implications for future research.	39; 42-46; 49
Other information				
Supplementary information	21	);V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	N/A
Funding	22	);V	Give the source of funding and the role of the funders for the present study.	N/A