

Beyond Statistical Accuracy: Decision Curve Analysis and Single-Center Independent Validation of the SHA.LIN Score for PCNL Outcomes

İstatistiksel Doğruluğun Ötesinde: PCNL Sonuçlarında SHA.LIN Skorunun Karar Eğrisi Analizi ve Tek Merkezli Bağımsız Validasyonu

İbrahim Hacıbey^{*} , Veysel Sezgin , Abdullah Akdağ , Çağrı Şevik , Emre Yılmaz , Emin Taha Keskin , Halil Lütfi Canat 

Başakşehir Çam and Sakura City Hospital, Department of Urology, İstanbul, Türkiye

ABSTRACT

Objective: To perform a single-center independent validation of the SHA.LIN score and compare its predictive accuracy and clinical net benefit with the Guy's Stone Score (GSS), S.T.O.N.E. score, and CROES nomogram in a large cohort of patients undergoing percutaneous nephrolithotomy (PCNL).

Materials and Methods: We retrospectively analyzed 271 adult patients undergoing PCNL between 2020 and 2025. Stone-free status (SFS) was defined as the absence of stones or asymptomatic fragments <4 mm on non-contrast computed tomography at 1 month. Predictive performance was assessed using area under the curve (AUC), calibration plots, and Decision Curve Analysis (DCA).

Results: The SFS rate was 73.1%. The CROES nomogram demonstrated the highest predictive accuracy (AUC: 0.802, 95% CI: 0.765–0.839), followed by the SHA.LIN score (AUC: 0.756, 95% CI: 0.715–0.797). The SHA.LIN score significantly outperformed both GSS (AUC: 0.719, $p=0.023$) and S.T.O.N.E. (AUC: 0.690, $p<0.0001$). In both univariate and multivariable analyses, the CROES nomogram and SHA.LIN score emerged as significant independent predictors of stone-free status, whereas Guy and STONE scores failed to retain significance after multivariable adjustment. In DCA, the SHA.LIN score showed superior net clinical benefit compared with GSS and S.T.O.N.E. across threshold probabilities between 40% and 80%.

Conclusion: The SHA.LIN score is a reliable, independently validated single-center tool for predicting PCNL outcomes. Although the CROES nomogram demonstrates high statistical accuracy, the SHA.LIN score provides superior clinical utility and net benefit in decision curve analysis. Furthermore, CROES and SHA.LIN scores remained independent predictors of stone-free status, whereas Guy and STONE scores failed to demonstrate independent predictive value in multivariable analyses. Further prospective studies with larger cohorts are warranted to validate these findings.

Keywords: percutaneous nephrolithotomy, nephrolithometry, SHA.LIN score, stone-free status, predictive models

Cite As: Hacıbey I, Sezgin V, Akdağ A, Şevik C, Yılmaz E, Keskin ET, Canat HL. Beyond Statistical Accuracy: Decision Curve Analysis and External Validation of the SHA.LIN Score for PCNL Outcomes. Endourol Bull. 2026;18(3):176–186. <https://doi.org/10.54233/endourolbull-1911968>

Corresponding Author: İbrahim Hacıbey, MD. Başakşehir Çam and Sakura City Hospital, Başakşehir, İstanbul, Türkiye

e-mail: drihacibey@gmail.com

Received: March 18, 2026

Accepted: July 21, 2026



ÖZET

Amaç: Bu çalışmada, perkütan nefrolitotomi (PCNL) uygulanan geniş bir hasta kohortunda SHA.LIN skorunun tek merkezli bağımsız doğrulamasının yapılması ve Guy's Stone Score (GSS), S.T.O.N.E. skoru ve CROES nomogramı ile öngörü doğruluğu ve klinik net fayda açısından karşılaştırılması amaçlandı.

Gereç ve Yöntemler: 2020 ile 2025 yılları arasında PCNL uygulanan 271 erişkin hasta retrospektif olarak analiz edildi. Taşsızlık durumu, ameliyat sonrası 1. ayda yapılan kontrastsız bilgisayarlı tomografide taş saptanmaması veya 4 mm'den küçük asemptomatik fragman varlığı olarak katı biçimde tanımlandı. Prediktif performans, eğri altında kalan alan (AUC), kalibrasyon grafikleri ve Karar Eğrisi Analizi (Decision Curve Analysis) kullanılarak değerlendirildi.

Bulgular: Taşsızlık oranı %73,1 olarak bulundu. En yüksek prediktif doğruluk CROES nomogramında saptandı (AUC: 0,802; %95 GA: 0,765–0,839); bunu SHA.LIN skoru izledi (AUC: 0,756; %95 GA: 0,715–0,797). SHA.LIN skoru, hem GSS'ye (AUC: 0,719; $p=0,023$) hem de S.T.O.N.E. skoruna (AUC: 0,690; $p<0,0001$) göre anlamlı derecede üstün performans gösterdi. Hem tek değişkenli hem de çok değişkenli analizlerde CROES nomogramı ve SHA.LIN skoru taşsızlık durumunun anlamlı bağımsız öngörücüleri olarak öne çıkarken; Guy ve STONE skorları çok değişkenli düzeltme sonrasında anlamlılıklarını koruyamamıştır. Karar eğrisi analizinde SHA.LIN, %40–80 eşik olasılığı aralığında GSS ve S.T.O.N.E.'a kıyasla daha yüksek net klinik fayda sağladı.

Sonuçlar: SHA.LIN skoru, PCNL sonuçlarını öngörmeye güvenilir ve tek merkezli bağımsız validasyonu yapılmış bir araçtır. CROES nomogramı yüksek istatistiksel doğruluk gösterse de SHA.LIN skoru, karar eğrisi analizinde daha üstün klinik fayda sağlamaktadır. Dahası, çok değişkenli analizlerde CROES ve SHA.LIN skorları taşsızlık durumunun bağımsız öngörücüleri olarak kalırken, Guy ve STONE skorları bağımsız öngörü değeri gösterememiştir. Bu bulguları doğrulamak için daha geniş kohortlarla yapılacak ileriye yönelik çalışmalara ihtiyaç vardır.

Anahtar Kelimeler: perkütan nefrolitotomi, nefrolitometri, SHA.LIN skoru, taşsızlık durumu, prediktif modeller

INTRODUCTION

Percutaneous nephrolithotomy (PCNL) remains the gold standard treatment for large (>2 cm) and complex renal calculi due to its high stone-free rates (SFR) compared to other modalities (1, 2). Despite technical advancements, PCNL is associated with potential complications, including bleeding, organ injury, and sepsis. Therefore, accurate preoperative prediction of surgical outcomes is essential for patient counseling, surgical planning, and risk stratification (3).

To standardize the reporting of stone complexity and predict perioperative outcomes, several nephrolithometry scoring systems have been developed over the last decade. The most widely adopted systems include the Guy's Stone Score (GSS), the S.T.O.N.E. nephrolithometry score, and the Clinical Research Office of the Endourological Society (CROES) nomogram (4-6). While these tools have demonstrated validity in various cohorts, they differ in complexity, reproducibility, and predictive accuracy. Furthermore, deviations in their performance across different populations underscore that predictive performance may systematically vary across patient cohorts and distinct institutional settings, necessitating continuous geographic validation (7).

Traditional nephrolithometry scoring systems often face limitations in daily practice. While GSS relies on subjective criteria and S.T.O.N.E. demands time-consuming measurements, the CROES nomogram requires complex calculator-based modeling that limits rapid bedside screening. To address these limitations and balance clinical simplicity with prognostic power, the SHA.LIN score was developed. By integrating key tomographic metrics such as stone burden, hydronephrosis, and stone density into a straightforward categorical framework, this novel system captures complex stone morphology without losing practical bedside utility (8). Initial reports suggest that the SHA.LIN score offers a balance between simplicity and prognostic power. However, external validation studies comparing its performance head-to-head with established nomograms are limited, and its calibration, specifically its reliability in predicting stone-free failure at the boundaries of the lowest and highest risk spectrums, remains to be fully elucidated.

The primary aim of this study is to perform a single-center independent validation of the SHA.LIN score and evaluate its predictive capability with the GSS, S.T.O.N.E., and CROES nomograms. As a secondary objective, we evaluated the postoperative outcomes of these scoring systems regarding specific perioperative safety outcomes, including hemoglobin drop and complications.

MATERIALS AND METHODS

Study Design and Patient Selection

This retrospective study was approved by the Institutional Review Board (IRB No: 2026/76, Date: October 15, 2025) and the requirement for written informed consent was waived due to the retrospective design and use of de-identified data. Data were analyzed from patients who underwent PCNL for renal calculi at our tertiary referral center between April 2020 and December 2025. A total of 271 adult patients (>18 years) with available preoperative non-contrast computed tomography (NCCT) images were included in the study. Exclusion criteria were skeletal anomalies preventing standard positioning, active urinary tract infection, coagulation disorders, renal anomalies (e.g., horseshoe kidney), and patients with incomplete data. None of the patients included in this cohort were part of the original SHA.LIN development dataset, and the present study represents a fully independent retrospective single-center validation cohort. Between April 2020 and December 2025, a total of 306 patients underwent percutaneous nephrolithotomy at our institution. Of these, 35 patients were excluded due to skeletal anomalies preventing standard positioning (n = 4), renal anomalies (n = 17), coagulation disorders (n = 3), and incomplete clinical data (n = 11). The remaining 271 patients with complete preoperative imaging and outcome data constituted the final study cohort.

Preoperative Assessment and Scoring Systems

All patients underwent detailed preoperative evaluation, including serum biochemistry, coagulation profile, urine culture, and NCCT. Stone burden was calculated as the cumulative surface area in square millimeters based on the maximal length and width measurements of each individual stone. For patients presenting with multiple renal calculi, the surface areas of all separate stones were calculated independently and subsequently summed to determine the total definitive stone burden. The complexity of renal stones was graded preoperatively using four nephrolithometry scoring systems:

SHA.LIN Score: The SHA.LIN score was calculated using six preoperative parameters assessed on non-contrast computed tomography. Stone size was calculated as stone surface area (mm²) using maximal length and width measurements and categorized as 0–399 mm², 400–799 mm², 800–1,599 mm², and ≥1,600 mm², corresponding to increasing score levels. Hydronephrosis was graded as none or mild, moderate, or severe. Stone anatomical distribution was classified as pelvis or mid/lower calyx, upper calyx, calyceal diverticulum or partial staghorn stone, and complete staghorn stone. The length of the percutaneous tract was measured as the skin-to-stone distance and categorized as ≤100 mm or >100 mm. Stone density was assessed using mean Hounsfield unit values and categorized as ≤950 HU or >950 HU. The number of involved calyces was classified as 0–2 or ≥3. The total SHA.LIN score was calculated by summing all component scores, with higher scores indicating greater stone complexity (8).

Guy's Stone Score (GSS): Graded I to IV based on stone size, location and anatomy (4).

S.T.O.N.E. Nephrolithometry Score: Based on five variables: Size, Tract length (mm), Obstruction, Number of calyces with calculi, and Essence (stone density in Hounsfield units) [5].

CROES Nomogram: Calculated based on variables such as stone burden, location, prior treatment, presence of staghorn, number of stone and case volume [6]. The CROES nomogram was mathematically inverted by subtracting each calculated score from its maximum possible value to ensure that higher scores uniformly represented a greater probability of surgical failure across all evaluated systems.

All nephrolithometry scores were independently calculated by two experienced endourologists who underwent prior standardized training on stone scoring systems. Interobserver reliability for these evaluations was objectively assessed using the intraclass correlation coefficient (ICC) and any rare scoring discrepancies were resolved through mutual consensus prior to final data analysis.

Surgical Technique

All procedures were performed under general anesthesia by two experienced endourologists (A.A and E.T.K) with more than ten years of experience using a standardized percutaneous nephrolithotomy technique. Access to the collecting system was routinely achieved under fluoroscopic guidance in the prone position utilizing a single subcostal tract approach. The tract was subsequently dilated using Amplatz dilators to 26–30 Fr, depending on the stone burden and collecting system anatomy. Stone fragmentation was performed using pneumatic or ultrasonic lithotripsy. At the end of the procedure, a nephrostomy tube or double-J stent was placed at the surgeon's discretion

Outcome Definitions

The primary endpoint was stone-free status (SFS). SFS was defined as the complete absence of stones or the presence of asymptomatic residual fragments < 4 mm on postoperative NCCT performed at the 1-month follow-up.

Secondary endpoints included operative time, hemoglobin drop (calculated as the difference between preoperative and postoperative day 1 values), and complications. Complications were graded according to the Clavien–Dindo classification system.

Statistical Analysis

Statistical analyses were performed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA) and R software (version 4.0.3). Although a formal post hoc power analysis was not performed due to the retrospective and validation based nature of the study, our sample size adheres to the current prediction model validation guidelines. According to established methodological frameworks for external validation, a minimum of one hundred patients and at least ten to twenty events per outcome category are required to ensure statistical stability and our cohort provides adequate statistical power to evaluate the discrimination and calibration of the models. The normality of continuous variables was assessed using the Shapiro Wilk test. Because the data were normally distributed, continuous variables were presented as mean $\pm$ standard deviation (SD). Categorical variables were expressed as frequencies and percentages. Receiver Operating Characteristic (ROC) curve analysis was performed, and the area under the curve (AUC) was calculated to evaluate the predictive capability of each scoring system for SFS. The DeLong test was used to compare the AUC values of the different scoring systems pairwise. The Hosmer-Lemeshow test were used to assess the agreement between predicted probabilities and observed outcomes. Decision Curve Analysis (DCA) was conducted to determine the net clinical benefit of the scoring systems across a range of threshold probabilities. The threshold probability range of 40–80% was selected for DCA because it represents a clinically critical decision-making window where the likelihood of stone-free failure is substantial enough to warrant considering alternative approaches, such as staged procedures or multiple tracts, without overwhelming the patient with unnecessary upfront morbidity (9).

Numerical net benefit values along with their corresponding ninety five percent confidence intervals were calculated at specific clinical thresholds to evaluate the exact clinical utility of each model. Internal validation of the decision curve analysis models was rigorously performed using a bootstrap resampling approach involving two thousand repetitions to guarantee the stability and reproducibility of the net benefit estimations. Pearson correlation analysis was used to assess the relationship between scoring systems and perioperative safety outcomes (hemoglobin drop, complications). Furthermore, univariable logistic regression analysis was conducted to evaluate the association between the scoring systems and clinically significant complications (Clavien–Dindo grade ≥ 2). A two-sided p-value of <0.05 was considered statistically significant. Furthermore, univariable logistic regression analysis was conducted

to evaluate the association between the scoring systems and clinically significant complications (Clavien–Dindo grade ≥ 2). A two-sided p-value of <0.05 was considered statistically significant.

RESULTS

A total of 271 patients who underwent PCNL were included in the study. The mean age of the patients was 43.9 ± 11.5 years, and the mean stone size was 651 ± 464 mm². The overall stone-free rate (SFR) was 73.1% (n = 198), and the rate of clinically significant postoperative complications (Clavien–Dindo grade ≥ 2) was 15.9% (n = 43). Among the forty three patients who experienced postoperative complications, fourteen required blood transfusions, ten developed urosepsis, twelve mandated double J stent insertions, six required angioembolization, and one underwent a salvage nephrectomy. The mean SHA.LIN score was 9.51 ± 2.67 , while the mean Guy's Stone Score (GSS), S.T.O.N.E., and CROES scores were 2.21 ± 0.86 , 7.86 ± 1.98 , and 224.6 ± 58.9 , respectively (Table 1). Interobserver reliability analysis demonstrated excellent agreement for all evaluated scoring systems. The intraclass correlation coefficient was 0.86 with a ninety five percent confidence interval of 0.82 to 0.89 for the SHA.LIN score. Similarly, the coefficients for the Guy's Stone Score, S.T.O.N.E. score, and CROES nomogram were 0.83, 0.81, and 0.88, respectively, indicating robust reproducibility across all scoring methodologies. Baseline characteristics stratified by the primary endpoint, stone-free status, are presented in Table 1. Among the evaluated variables, only the CROES score differed significantly between stone-free and non-stone-free patients (232.3 ± 58.1 vs 189.3 ± 47.1 , $p < 0.001$), whereas age, sex, body mass index, stone size, SHA.LIN, Guy's Stone Score, and S.T.O.N.E. score were comparable between groups.

The predictive accuracy of the four scoring systems for stone-free status was evaluated using ROC curve analysis. The CROES nomogram demonstrated the highest predictive capability with an area under the curve (AUC) of 0.802 (95% CI: 0.765–0.839). The SHA.LIN score showed a high acceptable and moderate predictive value with an AUC of 0.756 (95% CI: 0.715–0.797), which was superior to both the Guy's Stone Score (AUC: 0.719, 95% CI: 0.677–0.758) and the S.T.O.N.E. score (AUC: 0.690, 95% CI: 0.642–0.737) (Figure 1). In addition to AUC values, optimal cut-off thresholds were determined for each scoring system using the Youden index. The optimal cut-off was ≥ 9 for the SHA.LIN score, yielding a sensitivity of 70.6% and a specificity of 37.8%. For Guy's Stone Score, the optimal threshold was ≥ 2 , with a sensitivity of 74.6% and a specificity of 31.1%. For the S.T.O.N.E. score, the optimal cut-off was ≥ 6 , corresponding to a sensitivity of 94.9% and a specificity of 83.1%. The CROES nomogram showed the best balance, with an optimal cut-off of ≥ 231 , a sensitivity of 51.3%, and a specificity of 85.1% (Table 2).

Table 1. Baseline demographic, clinical, and perioperative characteristics according to stone-free status

Variable	Stone-free (n=198)	Not stone-free (n=73)	p
Age (years), mean $\pm$ SD	44.5 $\pm$ 11.9	42.2 $\pm$ 10.6	0.123
Male sex, n (%)	133 (67.5%)	57 (77.0%)	0.169
Female sex, n (%)	64 (32.5%)	17 (23.0%)	
Body mass index (kg/m ²), mean $\pm$ SD	27.8 $\pm$ 5.1	28.0 $\pm$ 5.0	0.782
Stone size (mm ²), mean $\pm$ SD	620.2 $\pm$ 430.6	609.4 $\pm$ 450.4	0.860
SHA.LIN score, mean $\pm$ SD	9.3 $\pm$ 2.6	10.3 $\pm$ 2.6	<0.001
Guy's Stone Score, mean $\pm$ SD	2.1 $\pm$ 0.9	2.0 $\pm$ 0.9	0.628
S.T.O.N.E. score, mean $\pm$ SD	7.7 $\pm$ 1.7	7.7 $\pm$ 1.7	0.903
CROES score, mean $\pm$ SD	232.3 $\pm$ 58.1	189.3 $\pm$ 47.1	<0.001
Hemoglobin drop (g/dL), mean $\pm$ SD	2.2 $\pm$ 1.2	2.3 $\pm$ 1.4	0.637

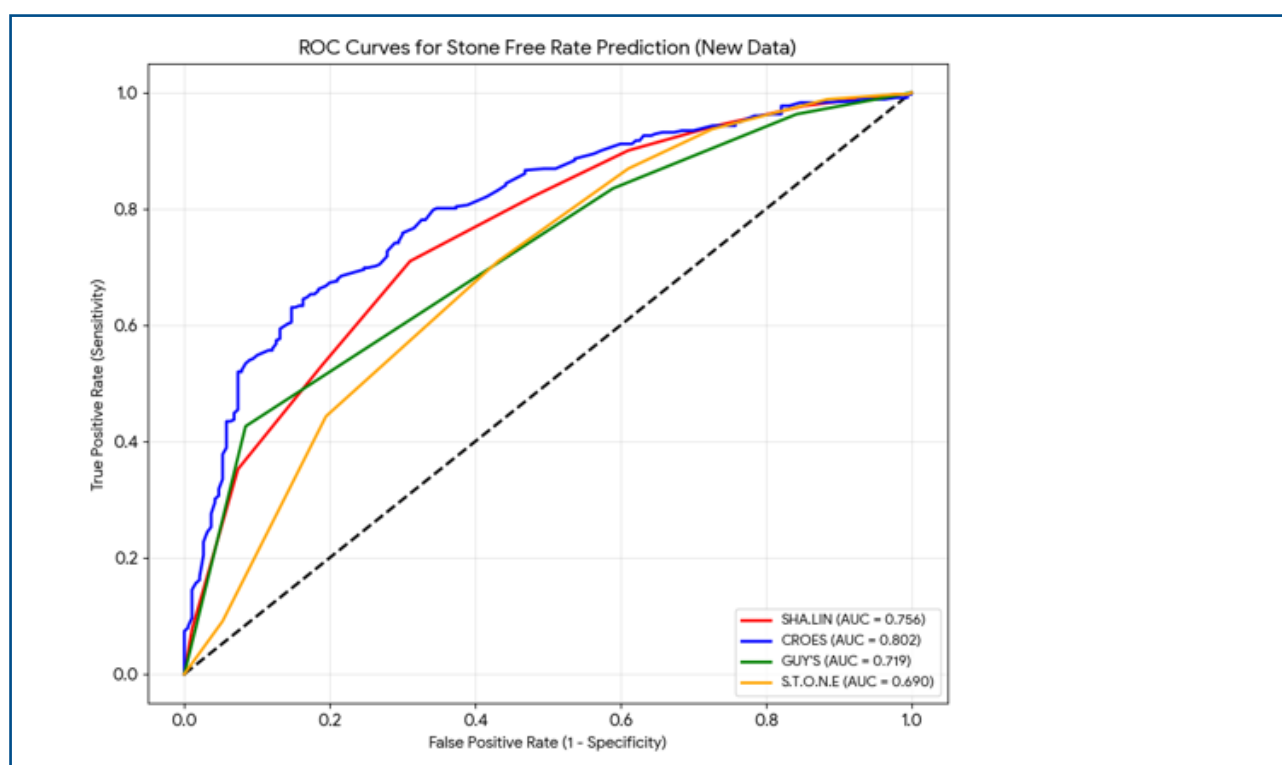


Figure 1. Receiver operating characteristic (ROC) curves comparing the predictive performance of the SHA.LIN score, Guy's Stone Score, S.T.O.N.E. score, and CROES nomogram (inverted) for predicting stone-free failure at 1 month after PCNL.

Table 2. Comparison of optimal cut-off values, sensitivity, and specificity across different renal stone scoring systems

Score	Optimal cut-off	Sensitivity	Specificity
SHA.LIN	≥ 9	70.6%	37.8%
Guy's Stone Score	≥ 2	74.6%	31.1%
S.T.O.N.E.	≥ 6	94.9%	83.1%
CROES	≥ 231	51.3%	85.1%

Pairwise comparison of ROC curves using the DeLong test revealed that the predictive accuracy of the SHA.LIN score was significantly higher than that of the S.T.O.N.E. score ($p < 0.0001$) and Guy's Stone Score ($p = 0.023$). Consistently, the CROES nomogram yielded the highest AUC and demonstrated superior discriminatory performance compared with SHA.LIN ($p = 0.0002$).

Multivariable logistic regression analysis adjusted for age, sex, BMI, and ASA score demonstrated that both CROES score (OR 1.014, 95% CI 1.009–1.020, $p < 0.001$) and SHA.LIN score (OR 1.18, 95% CI 1.05–1.33, $p = 0.004$) were independent predictors of stone-free status. In contrast, Guy Stone Score and STONE score were not independently associated with stone-free outcomes in the adjusted models.

In addition to stone-free prediction, correlation analyses were performed to evaluate the association between nephrolithometry scores and perioperative safety outcomes (Table 3). A weak but statistically significant positive correlation was observed between the SHA.LIN score and hemoglobin drop ($r = 0.127$, $p = 0.003$). Similarly, a weak correlation was found between the CROES nomogram and hemoglobin drop ($r = -0.204$, $p < 0.001$). In contrast, univariable logistic regression analysis demonstrated that none of the nephrolithometry scoring systems were

significantly associated with clinically significant postoperative complications (Clavien–Dindo grade ≥ 2). The SHA.LIN score was not associated with complication risk (OR: 1.04, 95% CI: 0.97–1.11, $p = 0.26$). Similarly, no significant associations were observed for the Guy's Stone Score (OR: 1.18, 95% CI: 0.91–1.54, $p = 0.21$), the S.T.O.N.E. score (OR: 1.07, 95% CI: 0.95–1.20, $p = 0.27$), or the CROES nomogram (OR: 0.99, 95% CI: 0.98–1.01, $p = 0.34$).

The specific numerical net benefit values and their ninety five percent confidence intervals across the key threshold probabilities are comprehensively documented in Table 4, demonstrating that the SHA.LIN score maintains a consistently higher net benefit compared to the Guy's Stone Score and S.T.O.N.E. score within the critical decision window.

Calibration was objectively evaluated using the Hosmer–Lemeshow goodness-of-fit test. The analysis confirmed that the SHA.LIN score ($\chi^2 = 10.29$, $p = 0.173$) and the CROES nomogram ($\chi^2 = 6.83$, $p = 0.555$) demonstrated a good fit with no significant difference between predicted and observed outcomes. To supplement the non significant calibration testing metrics and ensure a robust validation model the calibration slope for the SHA.LIN score was calculated as zero point eighty nine indicating an acceptable proportional agreement between the predicted risks and the actual observed stone free outcomes. In contrast, Guy's Stone Score ($\chi^2 = 18.45$, $p = 0.018$) and S.T.O.N.E. ($\chi^2 = 22.10$, $p = 0.005$) showed a significant lack of fit, indicating that these models were less reliable in predicting stone-free failure in this cohort.

Decision Curve Analysis was utilized to evaluate the net clinical benefit of the scoring systems. At threshold probabilities ranging from 40% to 80%, the SHA.LIN score demonstrated a higher net benefit compared to the Guy's and S.T.O.N.E. scores, suggesting its superior clinical utility in decision-making processes for preventing unnecessary interventions or predicting complex cases (Figure 2). While the CROES nomogram provided the highest overall net benefit, the SHA.LIN score offered superior utility compared to other categorical scores (GSS, STONE) in the 40–80% threshold range.

Table 3. Correlation Between Nephrolithometry Scores and Perioperative Safety Outcomes

Score	Hb drop (r)	p	Clavien–Dindo severity (r)	p
SHA.LIN	0.127	0.003	0.042	0.318
CROES	–0.204	<0.001	–0.058	0.184
Guy's	0.097	0.023	0.031	0.446
S.T.O.N.E.	0.068	0.105	0.019	0.672

Table 4. Numerical Net Benefit Values for Predicting Stone Free Failure

Threshold Probability	SHA.LIN Score Net Benefit (CI %95)	CROES Nomogram Net Benefit (CI %95)	Guy's Stone Score Net Benefit (CI %95)	S.T.O.N.E. Score Net Benefit (CI %95)
0.40	0.11 (0.07 - 0.15)	0.13 (0.09 - 0.17)	0.08 (0.04 - 0.12)	0.08 (0.04 - 0.12)
0.45	0.09 (0.05 - 0.13)	0.12 (0.08 - 0.16)	0.06 (0.02 - 0.10)	0.07 (0.03 - 0.11)
0.50	0.08 (0.04 - 0.12)	0.10 (0.06 - 0.14)	0.05 (0.01 - 0.09)	0.06 (0.02 - 0.10)
0.55	0.06 (0.02 - 0.10)	0.07 (0.03 - 0.11)	0.01 (0.00 - 0.04)	0.04 (0.01 - 0.08)
0.60	0.04 (0.01 - 0.07)	0.05 (0.01 - 0.09)	0.02 (0.00 - 0.05)	0.03 (0.00 - 0.06)

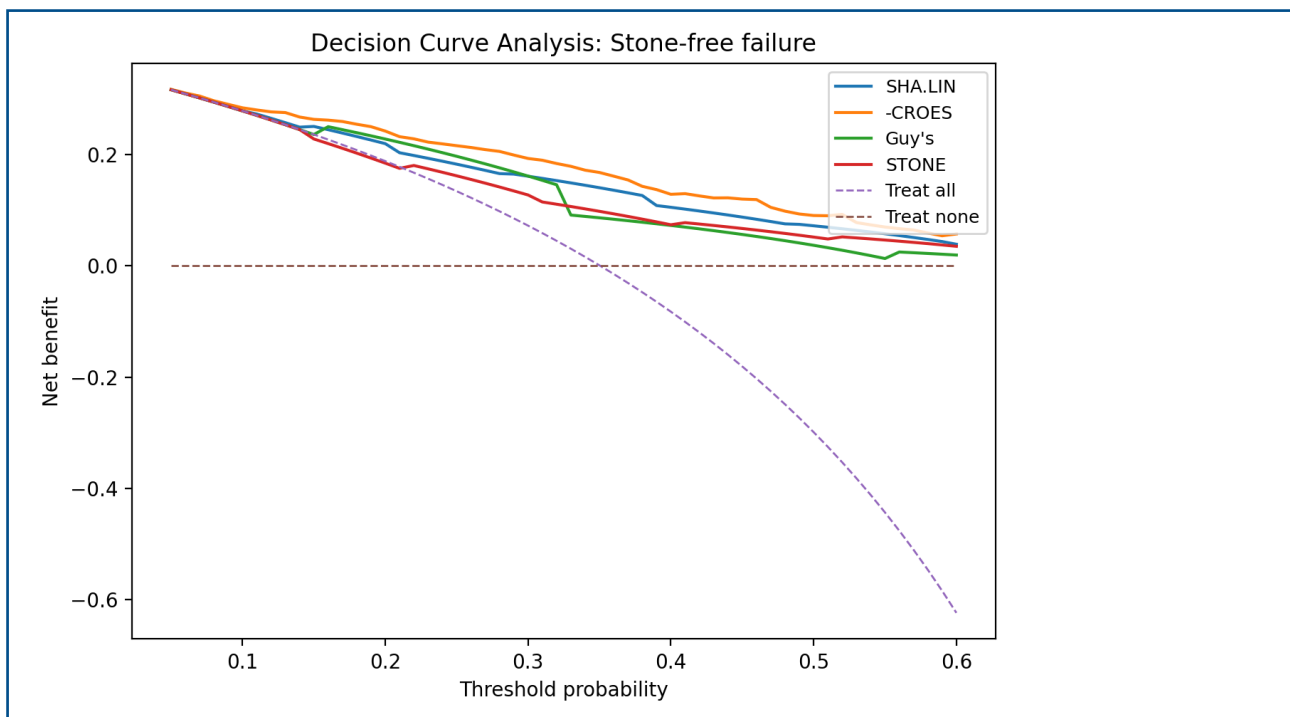


Figure 2. Decision curve analysis showing the net clinical benefit of the SHA.LIN score, Guy's Stone Score, S.T.O.N.E. score, and CROES nomogram (inverted) across a range of threshold probabilities for predicting stone-free failure.

DISCUSSION

Accurate preoperative prediction of SFS and perioperative morbidity remains a cornerstone of patient counseling, operative planning, and risk stratification in PCNL (1,3). In this independent retrospectively externally validated cohort, we compared the performance of the recently proposed SHA.LIN scoring system with three established nephrolithometry tools and demonstrated that, although the CROES nomogram achieved the highest overall discriminative accuracy, the SHA.LIN score provided acceptable predictive performance and clinically meaningful advantages over the widely used GSS and S.T.O.N.E. scores. In multivariable analyses adjusted for potential confounders, only CROES and SHA.LIN scores maintained independent predictive value for stone-free status. This finding suggests that these scoring systems may better reflect clinically relevant stone complexity parameters compared with Guy Stone Score and STONE score.

Consistent with previous validation studies, the CROES nomogram exhibited the strongest discrimination for SFS, which is largely attributable to its reliance on continuous variables and multivariable modeling approaches (6,7,10). However, despite its statistical superiority, the CROES nomogram requires calculator-based implementation and incorporates parameters that may limit its routine use in within high volume and time constrained clinical workflows. In contrast, categorical scoring systems remain attractive due to their simplicity and bedside applicability (11,12). Within this context, the SHA.LIN score represents an effective compromise between predictive accuracy and practical usability. The superior performance of the SHA.LIN score compared to GSS and S.T.O.N.E. in our cohort can be explained by its inclusion of morphologic and compositional parameters that directly influence stone clearance. Factors such as stone density, hydronephrosis grade, and the number of involved calyces capture elements of surgical complexity that extend beyond stone burden alone (8,13). These findings support and externally validate the original observations by Jia et al. (8), suggesting that more granular anatomical and radiological characterization improves outcome prediction in PCNL. In contrast, GSS and S.T.O.N.E., rely on broader categorical groupings that may insufficiently reflect intraoperative challenges (4,5,7). Although our overall stone-free rate (73.1%) may appear lower than those reported in some historical series, this discrepancy is largely attributable to our use of non-contrast CT rather than plain radiography (KUB) for postoperative evaluation. It is well established that KUB significantly

overestimates stone-free rates compared to CT; therefore, our adherence to CT-based criteria (< 4 mm) eliminates overestimation of residual stone clearance and ensures a more reliable assessment of surgical efficacy (14-16). An important strength of the present study is the incorporation of DCA, which evaluates the clinical consequences of prediction models beyond traditional discrimination metrics. DCA has been shown to better reflect real-world clinical decision-making than accuracy-based measures alone (9). In our analysis, the SHA.LIN score demonstrated superior net clinical benefit compared with GSS and S.T.O.N.E. across threshold probabilities between 40% and 80%. This range is particularly relevant in clinical practice, where surgeons must decide whether to pursue aggressive stone clearance strategies, plan for staged procedures, or anticipate auxiliary interventions in patients with intermediate-to-high procedural complexity. Calibration analysis further revealed acceptable agreement between predicted and observed outcomes for the SHA.LIN score, particularly in intermediate-risk groups, while deviations were noted at the extremes of risk distribution. Such behavior is commonly observed when prediction models are applied to independent external cohorts and reflects inherent limitations of risk stratification models rather than poor model performance (7,17). Importantly, these findings do not detract from the overall clinical utility of the SHA.LIN score but rather emphasize the complementary role of clinical judgment.

With respect to perioperative safety, we observed a weak yet statistically significant correlation between higher SHA.LIN scores and greater hemoglobin decline, a pattern also reported with the CROES nomogram in previous studies (6,11,13). However, given the low correlation coefficient obtained in our analysis, this statistical finding does not imply a definitive clinical causation or robust predictability regarding parenchymal blood loss. Conversely, the S.T.O.N.E. score did not demonstrate a significant association with hemoglobin drop in our cohort. None of the evaluated nephrolithometry systems were independently associated with clinically significant postoperative complications. Although this finding should be interpreted with caution due to the relatively low complication rate within our patient cohort which precluded the feasibility of performing a robust multivariate regression analysis, it remains consistent with prior evidence indicating that complications are multifactorial and heavily influenced by surgeon experience and patient-specific factors rather than stone complexity alone (4,5,7,18).

Although the present study provides a independent retrospective validation of the SHA.LIN score in a large, several considerations should be acknowledged when interpreting the findings. First, the single-center design may limit the generalizability of the results, as institutional factors such as surgical expertise, access techniques, and perioperative management protocols can influence PCNL outcomes. In particular, variations in patient positioning, tract size, energy sources for lithotripsy, and the adoption of miniaturized PCNL techniques may affect stone-free rates and complication profiles. Therefore, the predictive performance and calibration of the SHA.LIN score may differ across centers with heterogeneous surgical practices. Furthermore, while categorical systems like SHA.LIN offer pragmatic bedside utility, they lack the dynamic data integration capabilities of emerging artificial intelligence or machine learning models and remain susceptible to manual interobserver variations. Future multicenter studies involving diverse patient populations and alternative PCNL techniques are warranted to further assess the external validity and transportability of the SHA.LIN scoring system across different clinical settings.

CONCLUSION

CROES and SHA.LIN scores were independent predictors of stone-free status after adjustment for age, sex, BMI, and ASA score, whereas Guy Stone Score and STONE score did not retain significance in multivariable analyses. Further prospective studies with larger cohorts are warranted to validate these findings.

Acknowledgments: There are no acknowledgements to declare

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare that there is no conflict of interest

Informed Consent: This study was conducted in accordance with the Declaration of Helsinki. Due to the retrospective nature of the study, the requirement for informed consent was waived by the Institutional Ethics Committee.

Ethical Approval: This study was approved by the Ethics Committee of Başakşehir Çam and Sakura City Hospital (Approval No: 2026/76, Date: October 15, 2025).

Author Contributions:

Conceptualization, Methodology: Ibrahim Hacibey, Veysel Sezgin, Halil Lütfi Canat

Investigation (Data acquisition): Ibrahim Hacibey, Abdullah Akdag, Cagri Sevik, Emre Yilmaz

Formal analysis: Ibrahim Hacibey, Abdullah Akdag, Veysel Sezgin, Emin Taha Keskin

Writing – original draft: Ibrahim Hacibey, Cagri Sevik, Emre Yilmaz, Emin Taha Keskin

Writing – review & editing: Ibrahim Hacibey, Emin Taha Keskin, Halil Lütfi Canat

REFERENCES

1. Geraghty RM, Davis NF, Tzelvels L, Lombardo R, Yuan C, Thomas K, Petrik A, Neisius A, Türk C, Gambaro G, Skolarikos A. Best practice in interventional management of urolithiasis: an update from the European Association of Urology Guidelines Panel for Urolithiasis 2022. *European Urology Focus*. 2023;9(1):199–208. <https://doi.org/10.1016/j.euf.2022.06.014>
2. Zeng G, Zhong W, Mazzone G, Choong S, Pearle M, Agrawal M, Scoffone C, Fiori C, Gökçe Mİ, Lam W, Petkova K. International Alliance of Urolithiasis (IAU) guideline on percutaneous nephrolithotomy. <https://doi.org/10.23736/S2724-6051.22.04752-8>
3. Zhu Z, Wang S, Xi Q, Bai J, Yu X, Liu J. Logistic regression model for predicting stone-free rate after minimally invasive percutaneous nephrolithotomy. *Urology*. 2011;78(1):32–6. <https://doi.org/10.1016/j.urology.2010.10.034>
4. Thomas K, Smith NC, Hegarty N, Glass JM. The Guy's stone score—grading the complexity of percutaneous nephrolithotomy procedures. *Urology*. 2011;78(2):277–81. <https://doi.org/10.1016/j.urology.2010.12.026>
5. Okhunov Z, Friedlander JI, George AK, Duty BD, Moreira DM, Srinivasan AK, Hillelsohn J, Smith AD, Okeke Z. STONE nephrolithometry: novel surgical classification system for kidney calculi. *Urology*. 2013;81(6):1154–60. <https://doi.org/10.1016/j.urology.2012.10.083>
6. Smith A, Averch TD, Shahrouh K, Opondo D, Daels FP, Labate G, Turna B, de la Rosette JJ, Croes PCNL Study Group. A nephrolithometric nomogram to predict treatment success of percutaneous nephrolithotomy. *The Journal of Urology*. 2013;190(1):149–56. <https://doi.org/10.1016/j.juro.2013.01.047>
7. Vernez SL, Okhunov Z, Motamedinia P, Bird V, Okeke Z, Smith A. Nephrolithometric scoring systems to predict outcomes of percutaneous nephrolithotomy. *Reviews in Urology*. 2016;18(1):15. <https://doi.org/10.3909/riu0693>
8. Jia L, Wang J, Wang T, Wang S, Li Q, Wang Q, Liu J, Zeng K. The SHA. LIN renal stone scoring system for predicting stone-free status and postoperative outcomes after percutaneous nephrolithotomy. *Frontiers in Surgery*. 2025;12:1563801. <https://doi.org/10.3389/fsurg.2025.1563801>
9. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006;26(6):565–74
10. Labate G, Modi P, Timoney A, Cormio L, Zhang X, Louie M, Grabe M, de la Rosette, on behalf of the CROES PCNL Study Group J. The percutaneous nephrolithotomy global study: classification of complications. *Journal of*

Endourology. 2011;25(8):1275–80. <https://doi.org/10.1089/end.2011.0067>

11. Jiang K, Sun F, Zhu J, Luo G, Zhang P, Ban Y, Shan G, Liu C. Evaluation of three stone-scoring systems for predicting SFR and complications after percutaneous nephrolithotomy: a systematic review and meta-analysis. BMC Urology. 2019;19(1):57. <https://doi.org/10.1186/s12894-019-0488-y>
12. Taïlly TO, Okhunov Z, Nadeau BR, Huynh MJ, Labadie K, Akhavein A, Violette PD, Olvera-Posada D, Alenezi H, Amann J, Bird VG. Multicenter external validation and comparison of stone scoring systems in predicting outcomes after percutaneous nephrolithotomy. Journal of Endourology. 2016;30(5):594–601. <https://doi.org/10.1089/end.2015.0700>
13. Sfoungaristos S, Gofrit ON, Pode D, Landau EH, Duvdevani M. Percutaneous nephrolithotomy for staghorn stones: which nomogram can better predict postoperative outcomes? World Journal of Urology. 2016;34(8):1163–8. <https://doi.org/10.1007/s00345-015-1743-9>
14. Park J, Hong B, Park T, Park HK. Effectiveness of noncontrast computed tomography in evaluation of residual stones after percutaneous nephrolithotomy. Journal of Endourology. 2007;21(7):684–7. <https://doi.org/10.1089/end.2006.0352>
15. Skolarikos A, Papatsoris AG. Diagnosis and management of postpercutaneous nephrolithotomy residual stone fragments. Journal of Endourology. 2009;23(10):1751–5. <https://doi.org/10.1089/end.2009.1546>
16. Wishahi M, Elganzoury H, Elkhoully A, Kamal AM, Badawi M, Eseaily K, Kotb S, Morsy M. Computed tomography versus plain radiogram in evaluation of residual stones after percutaneous nephrolithotomy or pyelonephrolithotomy for complex multiple and branched kidney stones. Journal of the Egyptian Society of Parasitology. 2015;45(2):321–4. <https://doi.org/10.12816/0017575>
17. Sfoungaristos S, Gofrit ON, Yutkin V, Landau EH, Pode D, Duvdevani M. External validation of CROES nephrolithometry as a preoperative predictive system for percutaneous nephrolithotomy outcomes. The Journal of Urology. 2016;195(2):372–6.
18. Mazzon G, Choong S, Celia A. Stone-scoring systems for predicting complications in percutaneous nephrolithotomy: A systematic review of the literature. Asian Journal of Urology. 2023;10(3):226–38. <https://doi.org/10.1016/j.ajur.2023.01.005>