The New Journal of Urology
eISSN 3023-6940
doi: https://doi.org/10.33719/nju1388671

RESEARCH ARTICLE

Inter-Observer Reliability and Reproducibility of CROES, Guy's and STONE Nephrolithometry Scoring Systems for Predicting Percutaneous Nephrolithotomy Outcomes

Ali Ayrancı[1], Ufuk Cağlar[1], Hakan Çakır[2], Arda Meriç[1], Ufuk Can Aksu[1], Faruk Özgör[1], Ömer Sarılar[1]

[1] Department of Urology, Haseki Training and Research Hospital, Istanbul, Türkiye

[2] Department of Urology, Fulya Acibadem Hospital, Istanbul, Türkiye

Submitted: *2024-03-28*
Accepted: *2024-06-28*

Correspondence
Ali Ayranci, MD
Adress: Haseki Training and Research Hospital, Millet Street, Istanbul, Türkiye
E-mail: draliayranci@yahoo.com

ORCID

A.A.	0000-0003-3747-0869
U.Ç.	0000-0002-4832-9396
H.Ç.	0009-0003-7341-8360
A.M.	0000-0002-2611-2815
U.C.A.	0009-0003-4326-8197
F.Ö.	0000-0001-8712-7458
Ö.S.	0000-0002-1273-1084

Abstract

Objective: To assess inter-observer reliability and reproducibility of CROES, Guy's and S.T.O.N.E. nephrolithometry scoring systems (NSS).

Material and Methods: A total of 128 patients who underwent percutaneous nephrolithotomy (PNL) between January 2019 and January 2021 were included in the study. Calculation of the CROES, S.T.O.N.E, and Guy's NSSs was made by three independent urologists with different academic levels. These were; a very experienced (>500 PCNL cases) endourologist (Rater 1), a urologist who had just finished (>100 PCNL cases) their urology education (Rater 2) and a 3rd year urology resident who had never performed a PCNL operation (Rater 3). All were blinded to the procedure outcomes.

Results: An excellent correlation was found between three raters for Guy and S.T.O.N.E. scoring systems (kappa value 0.810-0.962). However, for the CROES score there is an excellent correlation between Rater 1 and Rater 2, but there were good correlations between Rater 1 vs Rater 3 and Rater 2 vs Rater 3 (kappa values 0.910 and 0.698-0.721 respectively). The highest correlation was between Rater 1 and Rater 2 for Guy score (kappa value 0.962) (Table 3). All intra-class correlations were statistically significant ($p<0.001$). The highest intra-class correlations were seen for the S.T.O.N.E. score (ICC: 0.980).

Conclusion: The present study revealed that all three NSS frequently used in current urology practice have reproducible and reliable results. Additionally, we believe that the application of CROES NSS by more experienced clinicians will be effective in obtaining clearer results.

Keywords: nomogram, scoring systems, percutaneous nephrolithotomy, surgery of renal stones

INTRODUCTION

Percutaneous nephrolithotomy (PNL) is an accepted treatment method for kidney stones greater than 2 cm (1). Success and complications may be affected by many factors including surgeon experience, renal anatomy, and complexity of renal stones. Outcomes of PNL have been reported to have wide ranges in the literature. Therefore, several nephrolithometry scoring systems (NSS) were developed for extensive patient counselling, surgical planning, and assessment of PNL results. Additionally, NSSs are used to

Cite; Ayranci A, Caglar U, Cakir H, Meriç A, Aksu UC, Ozgor F, Sarilar O. Inter-Observer Reliability and Reproducibility of CROES, Guy's and STONE Nephrolithometry Scoring Systems for Predicting Percutaneous Nephrolithotomy Outcomes. New J Urol. 2024;19(2):85-89. https://doi.org/10.33719/nju1388671

quantify the complexity of the stone in scientific papers (2).

Recently, three NSSs including the S.T.O.N.E. NSS, the Clinical Research Office of the Endourological Society (CROES) NSS and Guy's NSS are widely used in urology practice (3, 4, 5). After development of an NSS, its predictive accuracy is evaluated by internal and external validation. Although predictive accuracy seems to be the most important factor, simplicity, reproducibility, and achieving the same results with different clinicians are other important properties for an ideal nomogram. Optimally, scores achieved should be similar irrespective of the educational degree and level of expertise of the observer.

Although previous reports validated NSSs in predicting PNL outcomes, none of these studies compared the reproducibility and reliability of the 3 NSSs. In the present study, we evaluated CROES, Guy's and S.T.O.N.E. NSSs for reliability and reproducibility by analysing the scores calculated by 3 raters with different experience level.

MATERIAL AND METHODS

After approval from the Haseki Training and Research Hospital Clinical Research Ethics Committee (Approval Number:280-2023), we performed a retrospective study among patients who underwent conventional PNL from January 2019 to January 2021. Patients with missing data, patients < 18 years old, and patients who did not have pre-operative non-contrast abdominal computed tomography (NCCT) were excluded from the study. Evaluation of stone size, location, and density were evaluated by NCCT. All procedures were performed in the same manner and technique of PNL was described in detail previously. Stone-free status was assessed by NCCT 1-3 months later. Patients who had fragments not larger than 4 mm were considered stone free. Postoperative complications were categorized according to Clavien-Dindo (6).

Nephrolithometry Scoring System Assessment

Calculation of NSSs was made by three urologists with different experience levels (endourologist with >500 PNL, Rater 1; urologist with >100 PNL, Rater 2 and a 3rd year urology resident who had never performed a PNL, Rater 3). All were blinded to the procedure outcomes. CROES NSS (grade 1:0–100, grade 2:101–150, grade 3:151–200, and grade 4:201–350), S.T.O.N.E. NSS (scores between 5 and 13), and Guy's SS NSS (grade 1, 2, 3, 4) were analyzed (3,4,5). Case volume of the center: 500 cases per year.

Statistical Analysis

Statistical Package of Social Sciences for Windows (SPSS) version 20 was used. The compliance of data was evaluated by the Shapiro-Wilk test. Categorical variables were compared with Fisher's exact or Chi-square test. Sample *t* test was used for continuous parameters. Correlation analyses were done using Pearson's correlation coefficient. The Kappa value <0.20 reflects slight agreement, values of 0.21-0.40 are considered fair, 0.41-0.60 moderate, 0.61-0.80 good, and 0.81-1 indicates almost perfect agreement (7). Intra-class correlation was based on a two-way random effects model with type consistency. P <0.05 was considered statistically significant.

RESULTS

A total of 128 patients (86 males and 42 females) were included in the study. According to Rater 1, the mean stone size was 516.5 ± 370.6 mm^2, the mean stone-skin distance was 90.4 ± 24.2 mm and the mean Hounsfield unit (HU) was 1013.8 ± 301.0. The mean operation time and hospitalization time was 80.9 ± 32.7 minutes and 73.2 ± 45.3 hours respectively.

In total, 30 patients (23.4%) experienced complications. According to Clavien Dindo classification complication degree distribution was 6 patients with grade 1, 10 patients with grade 2, two patients with grade 3a and 12 patients with grade 3b. The stone-free status was achieved in 73.4% of patients (94 of 128 patients) (Table 1).

After all scoring systems were calculated by each rater, the Guy scores were 1.9 ± 0.9, 2.0 ± 0.9, and 2.1 ± 0.9, S.T.O.N.E. scores were 7.9 ± 1.4, 8.0 ± 1.4, and 8.8 ± 1.3, and CROES scores 202.9 ± 64.6, 203.7 ± 60.7, 173.1 ± 61.4, respectively according to raters (Table 2).

An excellent correlation was found between the three raters for Guy and S.T.O.N.E. scoring systems (kappa value 0.810-0.962). However, for the CROES score there was an excellent correlation between Rater 1 and Rater 2, but there were good correlations between Rater 1 vs Rater 3 and Rater 2 vs Rater 3 (kappa value 0.910 and 0.698-0.721 respectively). The highest correlation was between Rater 1 and Rater 2 for Guy score (kappa value 0.962) (Table 3). All intra-class correlations were statistically significant (p<0.001). The highest intra-class correlations were seen for the S.T.O.N.E. score (ICC: 0.980) (Table 4).

Table 1. Patient information

Number of Patients		128
Gender	**Female***	42(32.8%)
	Male*	86(67.2%)
Age (years)*		47.2 ± 14.6
BMI (kg/m²)*		27.4 ± 4.9
Operation side	**Right***	70(54.7%)
	Left*	58(45.3%)
Stone size (mm²)*α		516.5 ± 370.6
Stone - skin distance (mm)*α		90.4 ± 24.2
Hounsfield Unit*α		1013.8 ± 301.0
Operation time (minutes)*		80.9 ± 32.7
Hospitalization time (hours)*		73.2 ± 45.3
Stone free status		94 (73.4%)
Complications	**Total***	30 (23.4%)
	Grade 1*	6 (4.7%)
	Grade 2*	10 (7.8%)
	Grade 3a*	2 (1.6%)
	Grade 3b*	12 (9.3%)

*: mean±standard deviation or number (%)
α: According to rater 1
BMI: Body Mass Index

Table 2. Scoring systems according to raters

	Rater 1	**Rater 2**	**Rater 3**
CROES	202.9 ± 64.6	203.7 ± 60.7	173.1 ± 61.4
Guy	1.9 ± 0.9	2.0 ± 0.9	2.1 ± 0.9
S.T.O.N.E.	7.9 ± 1.4	8.0 ± 1.4	8.8 ± 1.3

CROES: Clinical Research Office of the Endourological Society, S.T.O.N.E.: stone size, tract length, degree of obstruction, number of involved calyces and stones' density

Table 3. Kappa correlation coefficient for all raters and scoring systems

Guy Score		
	Rater 2*	**Rater 3***
Rater 1*	0.962 (0.940-0.984)	0.810 (0.760-0.850)
Rater 2*		0.819 (0.775-0.863)
STONE Score		
	Rater 2*	**Rater 3***
Rater 1*	0.948 (0.923-0.973)	0.911 (0.878-0.944)
Rater 2*		0.910 (0.877-0.943)
CROES Score		
	Rater 2*	**Rater 3***
Rater 1*	0.910 (0.879 -0.941)	0.721 (0.671-0.771)
Rater 2*		0.698 (0.646-0.750)

*: 95% confidience
CROES: Clinical Research Office of the Endourological Society, S.T.O.N.E.: stone size, tract length, degree of obstruction, number of involved calyces and stones' density

Table 4. Intra-class correlation among all raters for scoring systems

	ICC	**95% Cl**	**p value**
Guy	0.978	0.970-0.984	***0.001***
STONE	0.980	0.974-0.986	***0.001***
CROES	0.964	0.951-0.973	***0.001***

Cl: confidence interval, ICC: Intra-class correlation
CROES: Clinical Research Office of the Endourological Society, S.T.O.N.E.: stone size, tract length, degree of obstruction, number of involved calyces and stones' density

DISCUSSION

Nomograms in surgical practice are usually used to predict complexity of disease and surgical outcomes; additionally they are used to determine the deviations from normality in internal medicine practice (8, 9, 10). The applicability and effectiveness of nomograms are frequently discussed and researched (11, 12). Researchers mostly focus on the ability of nomograms to predict outcomes; however, questioning of the compatibility and repeatability of the nomograms, and reliability between raters have not been clearly investigated (13). Nomograms can yield different results in terms of accuracy, but it is uncertain whether these differences are due to the nomograms themselves or the clinicians evaluating them. The question of who is the most suitable clinician to evaluate nomograms remains unanswered.

Three NSS are widely used in daily urology practice. In a recent meta-analysis, studies evaluating these three nomograms were examined and all three were stated to be suitable with equal power and accuracy in predicting stone-

free rates (14). However, another important situation is to compare whether these scoring systems always give the same results or not independently of the clinician applying them. In a recent study, the CROES nomogram was applied to their own patient population by 4 independent raters, and they stated that there was excellent agreement between the raters according to the nomogram scores. In the present study, an excellent correlation was determined between Guy's and S.T.O.N.E. NSSs. Experience affected the results most for the CROES NSS. When the correlation was evaluated for the CROES NSS, the results of 2 experienced raters were still perfectly compatible. However, the inexperienced clinician (rater 3) had a lower correlation score compared to both experienced raters. According to the intra-class correlation analyses, the highest correlation was seen for S.T.O.N.E. score and all three NSS achieved a statistical significance.

Analyzing the internal dynamics of these three NSS in detail revealed that there were differences in their natures. In the original article about CROES, stone burden, stone location and stone number are described with figures, but staghorn stone is not defined. The calculation of the score is done by the addition of 6 two-digit numbers marked on a scale. In addition, the size of the stone is obtained by a process that requires a calculator such as "$\text{width}_{max} \times \text{length}_{max} \times 0.785$". When "human error" is taken into account in the application of this score, it involves risks that will prevent obtaining the exact values (15). When the Guy's score is evaluated, the NSS is described with kidney illustrations and the final score is obtained by selecting one of the 4 categories. The lack of requirement for any mathematical operation makes the scoring system the simplest scoring system applied. Our third NSS of S.T.O.N.E. consists of 5 questions, and the answer to these questions comprise numbers from 1 to 4. Scores are obtained by adding these 5 single-digit numbers without the need for a calculator. In light of these explanations, the reason why the first two NSS have excellent correlation and the last one has good correlation between raters is due to the simple-complex nature of these NSSs. It is crucial to emphasize that for a nomogram to be clinically useful, its evaluations must demonstrate consistency across different observers. This consistency ensures that the outcomes are not influenced by subjective human factors, thus maintaining the reliability and validity of the tool. Our results indicate that the evaluations of all three nephrolithometry scoring systems (CROES, Guy's, and S.T.O.N.E.) exhibited high levels of agreement among raters with varying levels of experience. This high degree of similarity underscores the robustness and reproducibility of these nomograms, reinforcing their utility in clinical practice irrespective of the evaluator's expertise.

The retrospective nature of the study inherently introduces potential biases and limits the ability to establish causality. and relatively small patient number in the study could be considered limitations. Secondly, evaluating data for 128 patients according to 3 different nomograms in a short period of time of 1 week may have caused mental fatigue in the authors and affected their evaluation abilities.

CONCLUSION

We have two recommendations. First of all, this analysis could be made with a larger clinician population to achieve better conclusions and prevent mental fatigue of each clinician participating in the study. Secondly, we aimed to provide a guide for who is eligible to analyse nomograms but we did not evaluate whether the analysis could be made fully by artificial intelligence (AI) to prevent human error completely. We often see studies about AI in the field of radiology which may be subject of further studies in urology field (16).

The present study revealed that all three NSS have reproducible and reliable outcomes in prediction of PNL outcomes. Additionally, we found that the use of CROES NSS by more experienced clinicians will be effective in obtaining more clear results.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: None of the contributing authors have any conflict of interest, including specific financial interest and relationships and affiliations relevant to the subject matter of materials discussed in the manuscript.

Consent of Informed: The authors declare that written informed consent was obtained from the patients for publication of these reports.

Authors' Contributions: Conception and dizayn; Ayranci A, Data acquisition; Ayranci A, Cakir H, Data analysis and interpretation; Meriç A, Drafting the manuscript; Ayranci A, Cakir H, Critical revision of the manuscript for the content; Ayranci A, Ozgor F, Statistical analysis; Caglar U,

Supervision; Ozgor F, Sarilar O.

Ethics Committee: Haseki Training and Research Hospital Clinical Research Ethics Committee, Date: 01.02.2023, Protocol: 29-2023.

REFERENCES

1. Turk C, Neisius A, PEtrik A, Seitz C, Skolarikos A, Thomas K et al. EAU guidelines on urolithiasis. European Association of Urology. https://uroweb.org/guideline/prostate-cancer/ Edn. presented at the EAU Annual Congress Amsterdam 2020. EAU Guidelines Office, Arnhem, The Netherlands. ISBN 978-94-92671-07-3.
2. Sfoungaristos S, Mykoniatis I, Isid A, et al. Interobserver Reliability and Reproducibility of the Clinical Research Office of the Endourological Society Nomogram in Predicting Percutaneous Nephrolithotomy Results. Urology. 2016;97:56–60. https://doi.org/10.1016/j.urology.2016.07.014
3. Thomas K, Smith NC, Hegarty N, et al. The Guy's stone score--grading the complexity of percutaneous nephrolithotomy procedures. Urology. 2011;78:277-281. https://doi.org/10,1016/j.urology.2010.12.026
4. Okhunov Z, Friedlander JI, George AK, et al. S.T.O.N.E. nephrolithometry: novel surgical classification system for kidney calculi. Urology. 2013;81:1154- 1159. https://doi.org/10.1016/j.urology.2012.10.083
5. Smith A, Averch TD, Shahrour K, et al. A nephrolithometric nomogram to predict treatment success of percutaneous nephrolithotomy. J Urol. 2013;190:149-156. https://doi.org/10,1016/j.juro.2013.01.047
6. Clavien dindo: Clavien PA, Barkun J, de Oliveira ML et al (2009) The Clavien-Dindo classification of surgical complications: five-year experience. Ann Surg. 250:187-196 https://doi.org/10,1097/SLA.0b013e3181b13ca2
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. PMID: 843571
8. Cantinotti M, Giordano R, Clemente A, et al. Strengths and Limitations of Current Adult Nomograms for the Aorta Obtained by Noninvasive Cardiovascular Imaging. Echocardiography. 2016;33(7):1046-1068. https://doi.org/10,1111/echo.13232
9. Stojadinovic MM, Prelevic RI. External validation of existing nomograms predicting lymph node metastases in cystectomized patients. Int J Clin Oncol. 2015;20(1):164-170. https://doi.org/10.1007/s10147-014-0693-3
10. Onal B, Tansu N, Demirkesen O, et al. Nomogram and scoring system for predicting stone-free status after extracorporeal shock wave lithotripsy in children with urolithiasis. BJU Int. 2013;111(2):344-352. https://doi.org/10.1111/j.1464-410X.2012.11281.x
11. Ozgor F, Tosun M, Kayali Y, Savun M, Binbay M, Tepeler A. External Validation and Evaluation of Reliability and Validity of the Triple D Score to Predict Stone-Free Status After Extracorporeal Shockwave Lithotripsy. J Endourol. 2017;31(2):169-173. https://doi.org/10.1089/end2016.0721
12. Yanaral F, Ozgor F, Savun M, Sahan M, Sarilar O, Binbay M. Comparison of CROES, S.T.O.N.E, and Guy's scoring systems for the prediction of stone-free status and complication rates following percutaneous nephrolithotomy in patients with chronic kidney disease. Int Urol Nephrol. 2017;49(9):1569-1575. https://doi.org/10.1007/s11255-017-1631-x
13. Cantinotti, M., Scalese, M., Giordano, R. et al. A statistical comparison of reproducibility in current pediatric two-dimensional echocardiographic nomograms. Pediatr Res. 89, 579-590 (2021). https://doi.org/10.1038/s41390-020-0900-z
14. Jiang K, Sun F, Zhu J, et al. Evaluation of three stone-scoring systems for predicting SFR and complications after percutaneous nephrolithotomy: a systematic review and meta-analysis. BMC Urol. 2019;19(1):57. https://doi.org/10.1186/s12894-019-0488-y
15. Sameera, V., Bindra, A., & Rath, G. P. Human errors and their prevention in healthcare. Journal of anaesthesiology, clinical pharmacology, 2021;37(3):328-335. https://doi.org/10.4103/joacp.JOACP_364_19
16. Wang F, Gu X, Chen S, et al. Artificial intelligence system can achieve comparable results to experts for bone age assessment of Chinese children with abnormal growth and development. PeerJ. 2020;8:e8854. https://doi.org/10.7717/peerj.8854