HHS Public Access

Author manuscript

Ann Emerg Med. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

Ann Emerg Med. 2016 April ; 67(4): 423–432.e2. doi:10.1016/j.annemergmed.2015.08.019.

External Validation of the STONE Score, a Clinical Prediction Rule for Ureteral Stone: An Observational Multi-institutional Study

Ralph C. Wang, MD[*], **Robert M. Rodriguez, MD**, **Michelle Moghadassi, MPH**, **Vicki Noble, MD**, **John Bailitz, MD**, **Mike Mallin, MD**, **Jill Corbo, MD, RDMS**, **Tarina L. Kang, MD**, **Phillip Chu, PhD**, **Steve Shiboski, PhD**, and **Rebecca Smith-Bindman, MD**

Abstract

Study objective—The STONE score is a clinical decision rule that classifies patients with suspected nephrolithiasis into low-, moderate-, and high-score groups, with corresponding probabilities of ureteral stone. We evaluate the STONE score in a multi-institutional cohort compared with physician gestalt and hypothesize that it has a sufficiently high specificity to allow clinicians to defer computed tomography (CT) scan in patients with suspected nephrolithiasis.

Methods—We assessed the STONE score with data from a randomized trial for participants with suspected nephrolithiasis who enrolled at 9 emergency departments between October 2011 and February 2013. In accordance with STONE predictors, we categorized participants into low-, moderate-, or high-score groups. We determined the performance of the STONE score and physician gestalt for ureteral stone.

Results—Eight hundred forty-five participants were included for analysis; 331 (39%) had a ureteral stone. The global performance of the STONE score was superior to physician gestalt (area under the receiver operating characteristic curve=0.78 [95% confidence interval {CI} 0.74 to 0.81] versus 0.68 [95% CI 0.64 to 0.71]). The prevalence of ureteral stone on CT scan ranged from 14% (95% CI 9% to 19%) to 73% (95% CI 67% to 78%) in the low-, moderate-, and high-score groups. The sensitivity and specificity of a high score were 53% (95% CI 48% to 59%) and 87% (95% CI 84% to 90%), respectively.

Conclusion—The STONE score can successfully aggregate patients into low-, medium-, and high-risk groups and predicts ureteral stone with a higher specificity than physician gestalt.

[*]Corresponding Author. ralph.wang@ucsf.edu.

Author contributions: RCW, RMR, VN, and RS-B conceived the work. RCW, VN, JB, MM, JC, and TLK collected data. RCW and MM performed data cleaning. RCW, RMR, MM, JB, JC, TLK, PC, SS, and RS-B performed statistical analysis. RMR, VN, and RS-B participated in study design. RCW drafted the article, and all authors participated in its revision. All authors had full access to the data, take responsibility for the integrity of the data, and approved the article. RCW takes responsibility for the paper as a whole.

Presented at the Society of Academic Emergency Medicine conference, May 2015, San Diego, CA.

Clinical trial registration number: NCT01451931

The data were collected, results were analyzed, and the article was prepared without influence from funding agencies. The authors have followed the STROBE checklist in collecting and reporting their data, and elements of the checklist are incorporated into the article. This study was approved by the UCSF Committee on Human Research (institutional review board). Consent was obtained from patients for the parent randomized trial and data were completely deidentified for this secondary analysis. The raw data from this study are not currently available.

However, in its present form, the STONE score lacks sufficient accuracy to allow clinicians to defer CT scan for suspected ureteral stone.

INTRODUCTION

Background

Pain from a kidney stone is a common reason for US emergency department (ED) visits, accounting for more than 1 million visits annually.[1–3] Although most patients are discharged after an evaluation and symptomatic treatment, approximately 10% require inpatient admission.[1,4,5] Individuals who are unable to pass their stone may continue to experience pain, vomiting, and urinary symptoms, and ultimately require a urologic intervention.[6] The STONE score is a recently derived clinical prediction rule designed to aid clinicians to evaluate the risk of ureteral stone and important alternative diagnoses for patients with suspected nephrolithiasis.[6] The STONE score is calculated as a weighted sum of 5 categorical predictors; the points for each predictor are based on the estimated coefficients from a regression model constructed to predict the presence of a ureteral stone. Patients were classified into low-, moderate-, and high-score groups with corresponding outcome probabilities of ureteral stone and important alternative diagnoses. Patients with a high score had an 89% probability of ureteral stone and a 1.6% probability of alternative diagnosis; those with a low STONE score had a 9% probability of ureteral stone (the probability of alternative diagnosis was not reported in this group). In accordance with these outcome probabilities, the authors concluded that patients with a high STONE score could potentially receive ultrasonography, reduced-dose computed tomography (CT), or no further imaging. However, the authors did not report the sensitivity and specificity of the STONE score, which are important test characteristics of the decision rule, as opposed to the positive predictive value, which is heavily influenced by the prevalence of the outcome in the original study population.[7,8] A clinical decision rule that seeks to rule in ureteral stone should have an excellent specificity.[6–11]

Importance

Abdominal CT has become the most frequently used imaging test for suspected kidney stone because of its perceived superior diagnostic accuracy and ability to identify important alternative diagnoses, such as appendicitis and diverticulitis.[4,12–17] Despite a significant increase in the use of CT scans for patients with suspected kidney stone, there has been no demonstrable improvement in patient outcomes.[18–20] A recent national survey described a 10-fold increase in CT use during 1996 to 2007 for suspected kidney stone, without associated increases in kidney stone diagnoses, important alternative diagnoses, or hospitalization of kidney stone patients.[20] Furthermore, abdominal CT entails radiation exposure with attendant cancer risk, is associated with increased ED length of stay, and contributes to increasing annual care cost for acute nephrolithiasis, estimated in excess of $5 billion.[21–25] If the STONE score is found to identify patients with ureteral stone with sufficient accuracy without relying on further imaging, it could significantly improve the evaluation of patients with suspected nephrolithiasis.[7,9,26,27]

Goals of This Investigation

We sought to determine whether the STONE score could be used to safely decrease CT scan use in patients with suspected nephrolithiasis. Using data from a recently completed multicenter randomized trial comparing CT scan to ultrasonography for patients with suspected nephrolithiasis, we determined the discrimination, calibration, and test characteristics of the STONE score to predict ureteral stone. In addition, we compared the test characteristics of the STONE score to those of unstructured physician gestalt. We hypothesized that a high STONE score (10 to 13) would have sufficient specificity to diagnose ureteral stone and allow clinicians to defer CT scan in patients with suspected nephrolithiasis.

MATERIALS AND METHODS

Study Design and Setting

To evaluate the STONE score, we conducted a secondary analysis using data from a recently conducted randomized comparative effectiveness trial, the Study of Ultrasonography Versus Computed Tomography for Suspected Nephrolithiasis.[19] The randomized trial was conducted at 15 academic EDs across the United States between October 2011 and February 2013. Details of the participating EDs have been reported.[28] Briefly, the participating sites were academic EDs with emergency medicine residencies and emergency ultrasonography fellowships across the United States, with representation from a number of settings: urban, rural, university based, and safety net hospitals. The sites varied by size, annual census, and patient population served. This randomized trial was performed with institutional review board approval at each site and informed consent was obtained from all participants. This current study was performed with institutional review board approval at the University of California, San Francisco.

Selection of Participants

Adult participants with suspected kidney stones that required imaging (determined by an attending emergency physician) were randomly assigned to receive point-of-care ultrasonography, radiology ultrasonography, or CT as their initial imaging test. Patients were excluded from enrollment if they were pregnant, at high risk of an important alternative (non–kidney stone) diagnosis, had received a kidney transplant, required dialysis, had a known solitary kidney, or weighed more than 129 kg if men or 113 kg if women. The STONE score consists of 5 demographic and clinical variables collected during the ED visit: sex, race, nausea or vomiting, duration of pain symptoms, and hematuria on urine dipstick test. To closely model the validation study with the inclusion and outcome criteria of the original report, we restricted the analysis to sites that used urine dipstick testing for hematuria and to participants who underwent a CT scan during the index ED visit. Of the 2,759 total patients analyzed in the randomized trial, 845 participants had the data available for validation of the STONE score (Figure 1).

Methods of Measurement

Research coordinators used a standardized data collection form to collect detailed demographic, clinical, laboratory, and imaging data during the index ED visit. Before patient enrollment, research coordinators attended a 2-day meeting to receive training in study protocol, filling out forms, and data collection techniques. Additional weekly online meetings provided more training about data collection. Patients were directly interviewed for the subjective variables during the index ED visits. These data were recorded on paper forms and faxed to a data coordinating center, which provided immediate feedback for form completeness. Research coordinators were blinded to the study hypothesis. Dual assessments were not performed.

To calculate a STONE score for each participant in the validation cohort, we determined the presence of each of the 5 STONE score predictors (sex, timing [duration of time since symptom onset in hours], race [black versus nonblack], nausea/vomiting, and hematuria). Points were assigned when the predictor was present (male sex, 2 points; duration of pain <6 hours, 3 points; duration of pain 6 to 24 hours, 1 point; nonblack race, 3 points; nausea alone, 1 point; vomiting, 2 points; hematuria on urine dipstick testing, 3 points) after the initial report. The points received from individual predictors were summed to form a STONE score.

Treating physicians (either resident or attending) were asked to estimate the likelihood of kidney stone as the cause of the participants' symptoms. The physician could select from the following choices: 0% to 5%, 6% to 25%, 26% to 50%, 51% to 75%, and 76% to 100%. This question was asked before randomization and receipt of any imaging, and the physicians were blinded to the outcome of the study.

Outcome Measures

Ureteral stone was defined as the visualization of a kidney stone in the ureter (including stones at the ureteropelvic junction, ureter, and ureterovesicular junction) on CT. Important alternative diagnosis (such as pyelonephritis, malignancy, diverticulitis, pancreatitis, appendicitis, cholecystitis, pulmonary disorders, small bowel obstruction, and ovarian torsion) were defined with the same system of classification as used in the original study.[29] The presence of ureteral stone and alternative diagnosis was recorded during the index ED visit by trained research coordinators according to dictated CT reports. To assess the reproducibility of the ureteral stone and important alternative diagnosis outcomes for this study, detailed CT result dictations for each participant were obtained from 2 of the 9 sites (sites 6 and 8, including data from 103 participants, 12% of the validation cohort) and reviewed by one of the study authors (R.C.W.), who abstracted whether a ureteral stone or important alternative diagnosis was present and compared this with the data abstracted by the study coordinators in a blinded fashion. The interobserver κ agreement between study author (R.C.W.) and research coordinators for the ureteral stone outcome was 1.0 (perfect agreement) and 0.79 for important alternative diagnoses (good agreement).

Primary Data Analysis

One hundred twelve subjects (12%) had missing data for 1 or more variables (dipstick hematuria [n=85], duration of pain [n=9], race [n=9], nausea/vomiting [n=3], and ureteral stone [n=8]). We chose to focus our analysis of missing data on the urine dipstick variable. An analysis was performed to explore how results varied when values for the missing urine dipstick values were imputed. Missing values in the urine dipstick variable were replaced with imputed values. These imputed data were used in a multivariate logistic regression model, and the strength and direction of the associations between the STONE variables and ureteral stone were similar to those of the base model (Table E1, available online at http://www.annemergmed.com). Thus, we believe that it is acceptable to analyze data for participants with complete data.

We applied the STONE score to the validation cohort (n=845). Multivariate logistic regression was first performed to calculate odds ratios to determine the associations between the STONE score predictors with ureteral stone, specifying that the standard errors allow intrasite correlation. To examine the lack of association between the nonblack predictor and ureteral stone, we estimated the odds ratios of the STONE score predictors for ureteral stone in nonblack and black participants separately. The discrimination and calibration of the STONE score and physician gestalt were calculated with the area under the receiver operating characteristic curve (AUC) and Hosmer-Lemeshow goodness-of-fit test.

A score was calculated for each participant according to the observed predictor values. Participants were then categorized into low- (0 to 5), moderate- (6 to 9), and high-score (10 to 13) groups, and the prevalence of ureteral stones and important alternative diagnoses were determined for each group. The test characteristics of the STONE score were calculated, considering the high-score group (10 to 13 points) as a positive test result and ureteral stone on CT as a positive outcome. Similarly, the prevalence of ureteral stones and important alternative diagnoses were determined for each physician gestalt group. The test characteristics of physician gestalt were calculated, considering the 76% to 100% likelihood group as a positive test result and ureteral stone on CT as a positive outcome. An additional analysis was performed to assess whether the STONE score could be improved by omitting race (black versus nonblack). The AUC and test characteristics were calculated for this modified STONE score.

The uncertainty of the AUC and test characteristic estimates was summarized with exact binomial 95% confidence intervals (CIs). The interobserver reliability of the ureteral stone outcome measurement was determined with Cohen's κ. Stata (version 13; StataCorp, College Station, TX) was used to perform the statistical analysis.

RESULTS

Characteristics of Study Subjects

Of the 1,627 subjects who received CT scan in the randomized trial, 957 were enrolled at an ED that used urine dipstick testing. Of the 957, 112 were missing data (85 were missing urine dipstick) (Figure 1). The remaining 845 were included in the final analysis (mean age 40 years [range 18 to 75 years], 49% female patients, and 43% white) (Table 1). The

percentage of participants with a ureteral stone on CT was 39%, whereas the proportion of participants who had a significant alternative diagnosis was 5.3%. Eleven percent of participants required admission to an inpatient service from the ED.

Main Results

In the validation cohort, the overall direction of the associations between the STONE predictors and ureteral stone were similar to that of the original study, but the strength of the associations was attenuated; nonblack race was not significantly associated with ureteral stone (Table E1, available online at http://www.annemergmed.com). In nonblack participants, each of the STONE predictors was significantly associated with ureteral stone, but the duration of symptoms was not (Table E2, available online at http://www.annemergmed.com). The distribution of the duration of symptoms differed significantly in nonblack and black participants: 57% of nonblack participant presented with less than 24 hours of pain compared with 42% of black participants (P=.002).

Figure 2 graphically represents the receiver operating characteristic curves of the STONE score (as a numeric score 0 to 13) and physician gestalt (categorized into 0% to 5%, 6% to 25%, 26% to 50%, 51% to 75%, and 76% to 100% likelihood of ureteral stone). On inspection, the STONE score receiver operating characteristic curve appears to be closer to the left upper corner compared with that of physician gestalt. The main difference in the curves is the portion closest to the origin, suggesting that the STONE score specificity is superior to that of physician gestalt, whereas sensitivity is not much different. The AUC of the STONE score was 0.78 (95% CI 0.74 to 0.81) compared with that of physician gestalt, 0.68 (95% CI 0.64 to 0.71). It has been suggested that AUCs of 0.7 to 0.8 could be considered acceptable and those of 0.8 to 0.9 excellent.[30] The Hosmer-Lemeshow goodness-of-fit test of the STONE score resulted in a value of 8.5 (P=.40), indicating acceptable fit. Calibration of the STONE score was represented graphically by plotting observed versus predicted outcomes (Figure E1, available online at http://www.annemergmed.com).

Table 2 displays the prevalence of ureteral stone and alternative diagnoses in the groups with low, moderate, and high STONE score and the physician gestalt groups. The prevalence of ureteral stone ranged from 13.5% in the low-score group to 72.7% in the high-score group. The CIs of the estimates of prevalence do not overlap, indicating that the prevalence of ureteral stone was significantly different between the groups. The prevalence of important alternative diagnoses was 1.2% in the high-score group, but the upper limit of the 95% CI was 3.6%. The prevalence of ureteral stone and important alternative diagnosis in this cohort compared with that in the original validation study is graphically represented in Figure 3.

The prevalence of ureteral stone increased as the physician gestalt rating increased from 0% to 25%, to 76% to 100%. There was some overlap in the CIs of the estimates of ureteral stone prevalence, and the ranges of prevalence were more narrow when physician gestalt was used to categorize participants. Similar patterns were observed for the prevalence of important alternative diagnoses.

The test characteristics of the STONE score are displayed in Table 3. The authors of the original study suggested that patients with a high score (a STONE score of 10 to 13) could receive ultrasonography, reduced-dose CT scan, or, in some cases, no further imaging. Thus, we chose to report the test characteristics of the STONE score, considering a score of 10 to 13 as a positive test result. The sensitivity and specificity were 53% (95% CI 48% to 59%) and 87% (95% CI 84% to 90%), respectively. The STONE score positive likelihood ratio was 4.1, and negative likelihood ratio was 0.5. By considering a score of 11 to 13 as a positive test result, the specificity increased but the sensitivity decreased, and fewer participants would be considered as having a positive result. Conversely, by considering a STONE score of 5 to 13 as a positive test result, the sensitivity increased, but the specificity was drastically reduced.

The sensitivity and specificity of physician gestalt (considering a rating of 76% to 100% as a positive test result) were 62% (95% CI 56% to 67%) and 67% (95% CI 63% to 71%), respectively. At this 76% to 100% cutoff, the gestalt sensitivity is superior to the high STONE score, but with overlapping 95% CIs. The STONE score specificity is superior to physician gestalt at either cutoff.

We explored whether the STONE score could be improved by omitting the nonblack variable because it was not associated with ureteral stone in our cohort. The modified "STNE" score would then include sex, duration of symptoms, nausea or vomiting, and hematuria, and the score would range from 0 to 10 points. A comparison of the receiver operating characteristic curves can be found in Table E3 and Figure E2, available online at http://www.annemergmed.com. The AUC of the STONE score and the modified score were both 0.78, with overlapping 95% CIs. When defining a positive test result as a modified score of 8 to 10, the sensitivity and specificity were 42% (95% CI 37% to 48%) and 90% (87% to 92%), respectively, similar to the STONE score test characteristics.

LIMITATIONS

There are some limitations of this study. This was a secondary analysis of a randomized trial and thus is vulnerable to some methodological flaws. We did not validate the STONE score in all patients with suspected ureteral stone, but instead in patients who an attending emergency physician deemed should undergo CT scan imaging. This group likely represents those patients for whom attending emergency physicians have greater uncertainty about the diagnosis of kidney stone or alternative diagnosis compared with those who did not receive imaging. For example, patients deemed at very low risk or at very high risk for kidney stone (ie, classic renal colic in a patient with a history of kidney stones and hematuria) may have been less likely to be included. Also, to assemble a cohort with the available data for the STONE predictors and outcome, we restricted the study to subjects who received CT scan at sites that performed urine dipstick testing, which may have resulted in selection bias. Although this is a limitation of the present study, we believe that there is still value in assessing the STONE score in this population who required CT testing. Physicians' threshold for ordering CT scans for suspected kidney stone has decreased over time: emergency physician use of CT scan has increased 10-fold during the last 15 years, without a significant change in the diagnosis of kidney stone or rates of admission for kidney stone.[5]

The patients for whom there is greater diagnostic uncertainty should be the focus of a decision rule that functions to increase diagnostic certainty.

Another limitation is the lack of an assessment of the reliability of the predictor variables. Although we collected data prospectively in the randomized trial, dual assessments were not performed. Although sex, race, and hematuria on urine dipstick test are likely to have high interrater reliability, the presence of nausea or vomiting and the hours since the symptoms began require prospective evaluation. Finally, we sought to address potential measurement bias of the outcome and found that at 2 of the 9 sites, the interobserver agreement of the measurement of the ureteral stone outcome was perfect. We have no reason to believe that measurement bias would exist at the other 7 sites.

DISCUSSION

Using data from a large, randomized, comparative effectiveness trial, we evaluated the performance of the STONE score, a clinical decision rule derived to predict the presence of ureteral stone on CT scan. We compared the performance of the STONE score with that of physician gestalt, using several metrics of test performance, including discrimination, calibration, risk stratification, and test characteristics such as sensitivity, specificity, and likelihood ratios. The STONE score successfully categorizes patients into low-, moderate-, and high-risk groups, with corresponding probabilities of ureteral stone ranging from 13% to greater than 70%. The authors of the original study suggested that the STONE score could potentially be used to defer CT because patients with a high STONE score would be considered to have a ureteral stone and managed accordingly. A decision rule to identify patients with ureteral stone without need for further testing would require an excellent specificity and positive likelihood ratio. We found that the STONE score appears to have superior specificity compared with physician gestalt. However, a high STONE score was found to have a poor sensitivity (53%) and a moderate specificity (87%). In accordance with these findings, we believe that the STONE score does not have a sufficiently high specificity to defer CT scan without additional imaging.

We attempted to refine the STONE score to improve prediction, using the data from this validation cohort. We first explored the effect of changing the definition of a positive test result to a score of 11 to 13, which increased the specificity (92%; 95% CI 89% to 94%) and positive likelihood ratio, but worsened the sensitivity and negative likelihood ratio and decreased the proportion of patients identified with ureteral stone. We also considered modifying the STONE score to identify patients at very low risk for kidney stone, which would require an excellent sensitivity and negative likelihood ratio to defer CT. When the cutoff for a positive test result was decreased to a STONE score of 5, the sensitivity was improved but specificity was greatly reduced. The main problem with this approach is the presence of important alternative diagnoses in patients with suspected kidney stone. As in the original study, we found that the probability of ureteral stone was inversely related to the probability of an important alternative diagnosis. Thus, a low STONE score would not exclude important alternative diagnoses, which would likely be unacceptable to clinicians. Finally, we found that race (black versus nonblack) was not statistically associated with ureteral stone, perhaps because of the difference in the pattern of the duration of symptoms

between the nonblack and black participants. According to our sensitivity analysis, we did not find a significant difference in performance between the STONE score, including race and a modified score that omitted race. This suggests that the race predictor could be discarded in future studies of the STONE score. Ultimately, we could not improve the STONE score sufficiently to develop a decision rule that could provide a clear course of action with the available data.

There are some differences between the present study and the internal validation of the STONE score conducted by the authors of the original study. Our external validation cohort was larger and included participants from multiple institutions; it included a larger proportion of women and black participants. This is relevant to the STONE score because nonblack race and male sex are predictors in the score. Compared with the original internal validation, the STONE score performed with slightly lower discrimination (0.78 versus 0.82) in our cohort. Also, the prevalence of ureteral stone in the high-score group (ie, positive predictive value) in the original validation was 89%, and in this study it was 73%. This likely reflects the tendency of decision rules to perform less well in external cohorts compared with the population in which it was derived. Also, positive predictive values are known to vary with disease prevalence, and the overall prevalence was greater in the original study (56% versus 39%).[8]

Clinical decision rules that provide accurate outcome probabilities may be acceptable in clinical practice.[8] However, this type of rule does not clearly recommend a decision, and it is assumed that accurate predictions will improve clinical decisionmaking. Few decision rules of this type have undergone formal impact analysis, and clinicians do not know whether their use will actually improve patient outcomes compared with usual care. The Wells criteria were initially developed as a clinical decision rule that grouped patients into low, moderate, and high probabilities for pulmonary embolism. However, pulmonary embolism was not excluded in the low-risk group, and it was not clear how clinicians should interpret a low versus moderate score.[10] The Wells criteria were later combined with D-dimer testing to recommend a clear course of action, which allows clinicians to avoid CT imaging.[31] Similarly, more studies would be needed to combine the STONE score with additional testing (such as ultrasonography) to allow clinicians to defer CT imaging.

In summary, the STONE score successfully aggregated patients into low-, moderate-, and high-risk groups for ureteral stone. Also, it was superior to physician gestalt for predicting ureteral stone. However, the specificity of a high STONE score was modest and likely not sufficient to provide a clear course of action. These observations suggest that further development of the STONE score is needed to produce a successful decision rule that would allow clinicians to defer CT scan.[7,10,32] Alternatively, a new decision instrument could be derived and validated to improve the evaluation of ureteral stone by allowing clinicians to defer CT imaging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge Amy Markowitz, JD, Michael Kohn, MD, MPP, and Chris Moore, MD, for their excellent suggestions on the article, and Jersey Neilson, MAS, for help with data collection and management.

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist and provided the following details: This study was supported by funding from Agency for Healthcare Research and Quality (AHRQ) grant K08 HS02181 and National Center for Advancing Translational Sciences (NCATS) grant 8 KL2 TR000143-08. The Study of Ultrasonography Versus Computed Tomography for Suspected Nephrolithiasis was supported by the American Recovery and Reinvestment Act of 2009 through AHRQ grant R01HS019312.

References

1. Foster G, Stocks C, Borofsky MS. Statistical brief #139. 2012:1–10.
2. Brown J. Diagnostic and treatment patterns for renal colic in US emergency departments. Int Urol Nephrol. 2006; 38:87–92. [PubMed: 16502058]
3. Pearle MS, Pierce HL, Miller GL, et al. Optimal method of urgent decompression of the collecting system for obstruction and infection due to ureteral calculi. J Urol. 1998; 160:1260–1264. [PubMed: 9751331]
4. Fwu CW, Eggers PW, Kimmel PL, et al. Emergency department visits, use of imaging, and drugs for urolithiasis have increased in the United States. Kidney Int. 2013; 83:479–486. [PubMed: 23283137]
5. Westphalen AC, Hsia RY, Maselli JH, et al. Radiological imaging of patients with suspected urinary tract stones: national trends, diagnoses, and predictors. Acad Emerg Med. 2011; 18:699–707. [PubMed: 21762233]
6. Moore CL, Bomann S, Daniels B, et al. Derivation and validation of a clinical prediction rule for uncomplicated ureteral stone—the STONE score: retrospective and prospective observational cohort studies. BMJ. 2014; 348:g2191. [PubMed: 24671981]
7. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med. 1999; 33:437–447. [PubMed: 10092723]
8. Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. Ann Emerg Med. 2014; 64:286–291. [PubMed: 24530108]
9. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997; 277:488–494. [PubMed: 9020274]
10. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006; 144:201–209. [PubMed: 16461965]
11. Toll DB, Janssen KJM, Vergouwe Y, et al. Validation, updating and impactof clinical prediction rules: a review. J Clin Epidemiol. 2008; 61:1085–1094. [PubMed: 19208371]
12. Smith RC, Rosenfield AT, Choe KA, et al. Acute flank pain: comparison of non-contrast-enhanced CT and intravenous urography. Radiology. 1995; 194:789–794. [PubMed: 7862980]
13. Smith RC, Verga M, McCarthy S, et al. Diagnosis of acute flank pain: value of unenhanced helical CT. AJR Am J Roentgenol. 1996; 166:97–101. [PubMed: 8571915]
14. Teichman JMH. Clinical practice. Acute renal colic from ureteral calculus. N Engl J Med. 2004; 350:684–693. [PubMed: 14960744]
15. Fowler KA, Locken JA, Duchesne JH, et al. US for detecting renal calculi with nonenhanced CT as a reference standard. Radiology. 2002; 222:109–113. [PubMed: 11756713]
16. Mandavia DP, Pregerson B, Henderson SO. Ultrasonography of flank pain in the emergency department: renal cell carcinoma as a diagnostic concern. J Emerg Med. 2000; 18:83–86. [PubMed: 10645844]
17. Gottlieb RH, La TC, Erturk EN, et al. CT in detecting urinary tract calculi: influence on patient imaging and clinical outcomes. Radiology. 2002; 225:441–449. [PubMed: 12409578]

18. Ripollés T, Agramunt M, Errando J, et al. Suspected ureteral colic: plain film and sonography vs unenhanced helical CT. A prospective study in 66 patients. Eur Radiol. 2004; 14:129–136. [PubMed: 12819916]
19. Smith-Bindman R, Aubin C, Bailitz J, et al. Ultrasonography versus computed tomography for suspected nephrolithiasis. N Engl J Med. 2014; 371:1100–1110. [PubMed: 25229916]
20. Westphalen AC, Hsia RY, Maselli JH, Wang R, Gonzales R. Radiological Imaging of Patients with Suspected Urinary Tract Stones: National Trends, Diagnoses, and Predictors. Acad Emerg Med. 2011; 18:699–707. [PubMed: 21762233]
21. Kartal M, Eray O, Erdogru T, et al. Prospective validation of a current algorithm including bedside US performed by emergency physicians for patients with acute flank pain suspected for renal colic. Emerg Med J. 2006; 23:341–344. [PubMed: 16627832]
22. Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010. JAMA. 2012; 307:2400–2409. [PubMed: 22692172]
23. Smith-Bindman R, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. Health Aff (Millwood). 2008; 27:1491–1502. [PubMed: 18997204]
24. Kocher KE, Meurer WJ, Desmond JS, et al. Effect of testing and treatment on emergency department length of stay using a national database. Acad Emerg Med. 2012; 19:525–534. [PubMed: 22594356]
25. Gardner RL, Sarkar U, Maselli JH, et al. Factors associated with longer ED lengths of stay. Am J Emerg Med. 2007; 25:643–650. [PubMed: 17606089]
26. Auleley GR, Ravaud P, Giraudeau B, et al. Implementation of the Ottawa Ankle Rules in France. A multicenter randomized controlled trial. JAMA. 1997; 277:1935–1939. [PubMed: 9200633]
27. Blackmore CC. Clinical prediction rules in trauma imaging: who, how, and why? Radiology. 2005; 235:371–374. [PubMed: 15858080]
28. Valencia V, Moghadassi M, Kriesel DR, et al. Study of Tomography of Nephrolithiasis Evaluation (STONE): methodology, approach and rationale. Contemp Clin Trials. 2014; 38:92–101. [PubMed: 24721483]
29. Moore CL, Daniels B, Singh D, et al. Prevalence and clinical importance of alternative causes of symptoms using a renal colic computed tomography protocol in patients with flank or back pain and absence of pyuria. Acad Emerg Med. 2013; 20:470–478. [PubMed: 23672361]
30. Hosmer, DW., Jr; Lemeshow, S. Applied Logistic Regression. 2. New York, NY: John Wiley & Sons; 2000.
31. van Belle A, Büller HR, Huisman MV, et al. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. JAMA. 2006; 295:172–179. [PubMed: 16403929]
32. McGinn TG, Guyatt GH, Wyer PC, et al. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA. 2000; 284:79–84. [PubMed: 10872017]

Editor's Capsule Summary

What is already known on this topic

The STONE score is a clinical decision rule to risk-stratify urolithiasis.

What question this study addressed

Can the STONE score be used to rule in stones such that computed tomography (CT) scanning is unnecessary?

What this study adds to our knowledge

In this validation study of 845 adults receiving CT scanning for suspected urolithiasis, using a high-risk score rather than CT to rule in urolithiasis identified 53% of stones while falsely suggesting stones in 13% of patients without calculi. Furthermore, one of the score's 5 core elements failed to predict urolithiasis.

How this is relevant to clinical practice

This independent assessment found the STONE score to be an inaccurate tool to defer CT scanning and identified one of its core elements as invalid.

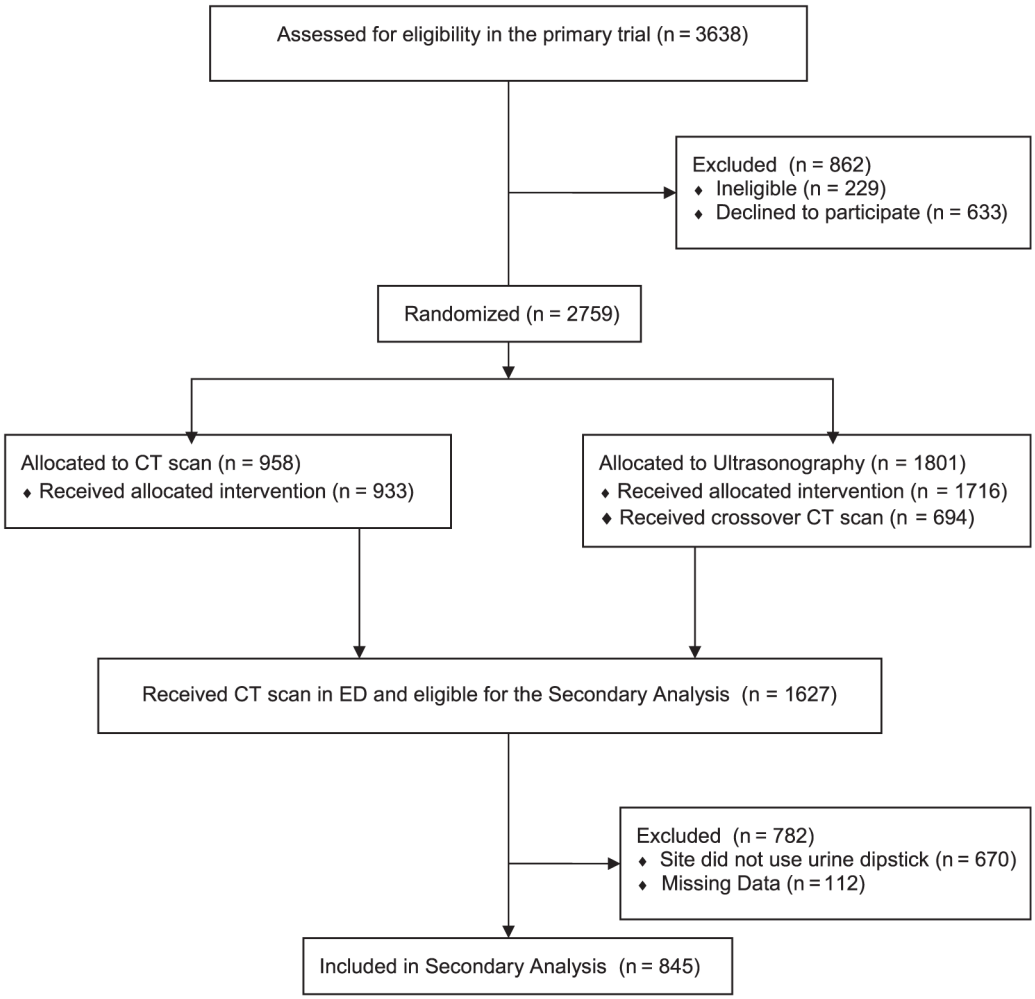

Figure 1.
Patient flow diagram.

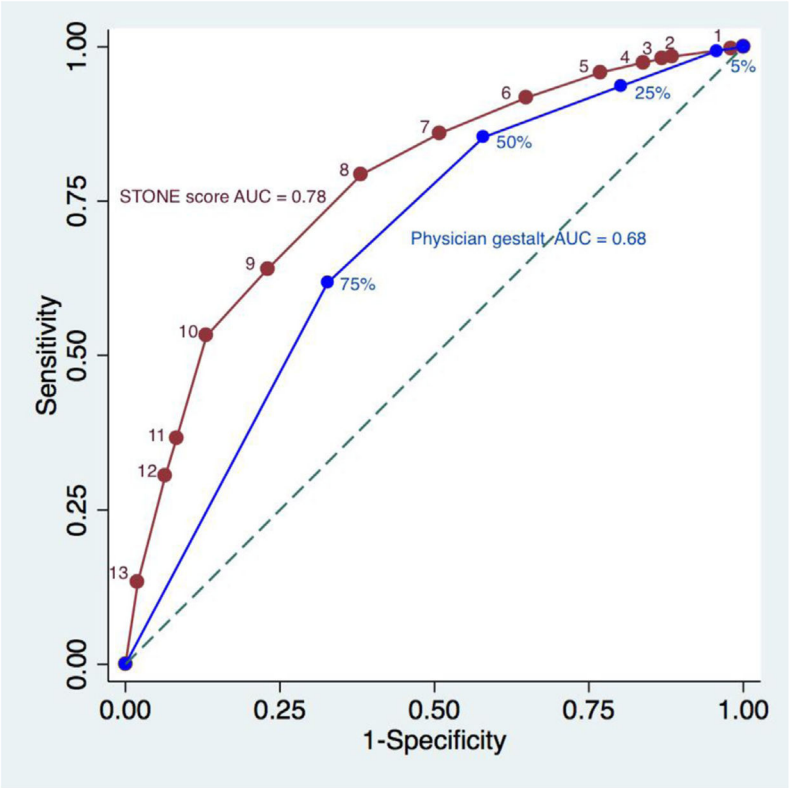

Figure 2.

Receiver operating characteristic curves of the STONE score and physician gestalt. Numbers indicate the STONE score and physician gestalt cutoffs.

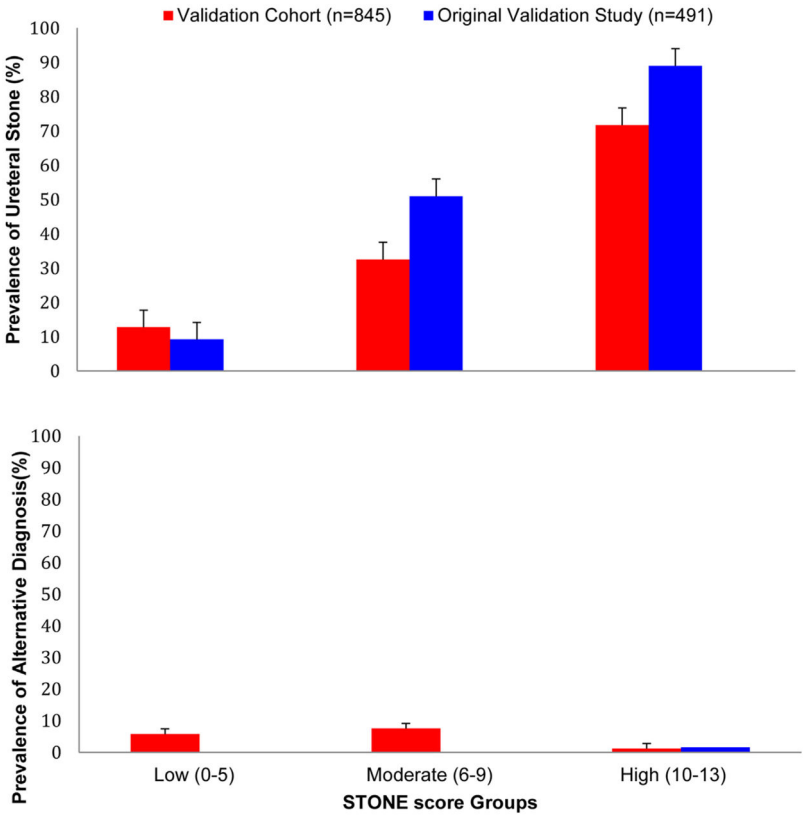

Figure 3.

Prevalence of ureteral stone and important alternative diagnoses by STONE score versus original validation study.[6] Error bars represent 95% CIs.

Table 1

Baseline characteristics of participants in this validation cohort versus the original validation study.[6]

Characteristics	This Validation Cohort (n=845), Frequency (%)	Original Validation Study* (n=491), Frequency (%)
Age (SD), y	40 (13)	46 (15)
Male sex	413 (49)	273 (56)
Race		
White	359 (43)	411 (84)
Black	169 (20)	57 (12)
Hispanic	235 (28)	†
Asian	45 (5)	†
Mixed	27 (3)	All other
Native American Indian	8 (1)	23 (5)
Pacific Islander	2 (0.2)	†
Radiology findings		
Ureteral stone	331 (39.2)	274 (56)
Alternative diagnosis	45 (5.3)	18 (3.7)
Disposition		
Admission to hospital	95 (11)	52 (11)

*Data from the original study describe the internal validation cohort.[6]

†Data from the initial study are not available.

Table 2

Prevalence of ureteral stone and important alternative diagnoses compared with physician gestalt in this validation cohort and in the original validation study.

		Prevalence (%) [95% CI]	
Risk Score	**Frequency (% of Cohort)**	**Ureteral Stone**	**Alternate Diagnosis**
STONE score in this validation cohort			
High	242/845 (28.6)	176/242 (72.7) [66.7–78.2]	3/242 (1.2) [0.3–3.6]
Moderate	395/845 (46.7)	127/395 (32.2) [27.6–37.0]	30/395 (7.6) [5.2–10.7]
Low	208/845 (24.6)	28/208 (13.5) [9.1–18.9]	12/208 (5.8) [3.0–10.0]
Physician gestalt in this validation cohort, %			
76–100	356/809 (44.0)	194/356 (54.5) [49.2–59.8]	10/356 (2.8) [1.4–5.1]
51–75	199/809 (24.6)	74/199 (37.2) [30.5–44.3]	9/199 (4.5) [2.1–8.4]
26–50	136/809 (16.8)	26/136 (19.1) [12.9–26.7]	12/136 (8.8) [4.6–14.9]
0–25	118/809 (14.6)	20/118 (17.0) [10.7–25.0]	9/118 (7.6) [3.5–14.0]
STONE score in the original validation study			
High	185/491 (37.7)	164/185 (88.6) [83.1–92.8]	1.6
Moderate	230/491 (46.8)	118/230 (51.3) [44.6–57.9]	*
Low	76/491 (15.5)	7/76 (9.2) [3.8–18.0]	*

* Information not available.

Table 3

Test characteristics of the STONE score and physician gestalt for ureteral stone.

Risk Score	Sensitivity, % (95% CI)	Specificity, % (95% CI)	DLR, % (95% CI)	−LR, % (95% CI)
STONE score (10–13)	53 (48–59)	87 (84–90)	4.1 (3.2–5.3)	0.5 (0.5–0.6)
STONE score (11–13)	37 (32–42)	92 (89–94)	4.6 (3.3–6.3)	0.7 (0.6–0.7)
STONE score (5–13)	96 (93–98)	23 (19–27)	1.2 (1.2–1.3)	0.2 (0.1–0.3)
Gestalt (76%–100%)	62 (56–67)	67 (63–71)	1.9 (1.6–2.2)	0.6 (0.5–0.7)
Gestalt (50%–100%)	85 (81–89)	42 (38–47)	1.5 (1.4–1.6)	0.3 (0.3–0.4)