Critical Review Form
Clinical Prediction or Decision Rule

Derivation and validation of a clinical prediction rule for uncomplicated ureteral stone--the STONE score: retrospective and prospective observational cohort studies. BMJ. 2014 Mar 26;348:g2191.

Objectives: To derive and validate "an clinical prediction score for ureteral stones that cause symptoms, identifying patients with either a very high or very low probability of having an uncomplicated ureteral stone." (p. 2)

Methods: This study involved a retrospective derivation and prospective validation of a clinical prediction score in 2 emergency departments (EDs), one a large academic ED (Yale New Haven) and the other a freestanding ED associated with Yale New Haven (Shoreline Medical Center). All patients aged 18 years and older who underwent a CT "flank pain protocol" in either ED between April 2005 and November 2010 were eligible for the derivation cohort. Exclusion criteria included absence of back or flank pain, trauma, evidence of infection, active malignancy, known renal disease (creatinine > 1.5 mg/dL), or previous lithotripsy or ureteral stent.

Out of over 5383 subjects undergoing CT, 1040 eligible subjects with complete records were randomly selected for inclusion. Five physicians from 3 specialties (emergency medicine, internal medicine, and urology) identified an a priori list of factors potentially predictive of a ureteral stone. The presence or absence of these factors was abstracted from the medical record using a standardized form by personnel blinded to CT results. A subset of 50 records were randomly selected and blindly reviewed to assess inter-rater reliability for each factor through calculation of kappa values; factors with a κ of < 0.60 were not eligible for inclusion in the prediction rule. The results of the dictated CT reports were blindly abstracted: a kidney stone was felt to be the cause of the patient's symptoms if it lay between the renal pelvis and the ureterovesical junction. Inter-rater reliability of CT scan results was performed on a random selection of records as well.

Multivariate logistic regression was employed to create the best model, i.e. the model with the lowest misclassification rate and highest area under the curve (AUC). The five factors chosen and their point assignments are shown in Table 1, with a score ranging from 0-13. This STONE score had a misclassification rate of 0.23 (95% CI 0.22-0.23) and an AUC of 0.82 (0.74-0.90). After construction of the point system, the authors calculated scores equivalent to low (about 10%, STONE score 0-5); moderate (about 50%, STONE score 6-9); and high (about 90%, STONE score 10-13) probability of ureteral stone in the

derivation set. Inter-rater reliability for CT results was excellent (κ = 0.75-0.80)

Table 1. STONE score components

Factor	Points
Sex	
- Female	0
- Male	2
Timing	
- > 24 hours	0
- 6-24 hours	1
- < 6 hours	3
Race	
- Black	0
- Non-black	3
Nausea and vomiting	
- None	0
- Nausea alone	1
- Vomiting	2
Hematuria (on urine dipstick)	
- Absent	0
- Present	3

Between May 25, 2011 and January 24, 2013, consecutive patients presenting during defined periods in whom the physician planned to obtain a CT to evaluate for kidney stone were approached for enrollment in the validation cohort. Data on patients enrolled was collected prospectively by research associated blinded to the STONE score and the CT results. Patients were assigned point values from 0-13 based on the derived STONE score and were classified based on these scores as low, moderate, or high risk. CT results were categorized by associates blinded to the clinical factors (except the laterality of the pain).

A total of 491 patients were enrolled in the validation cohort. For this group, the STONE score had an AUC of 0.79 (95% CI 0.76-0.83). Acutely important alternative causes of pain were found in 3.7% of this group overall; in the high probability group, representing 37.7% of the cohort, an acutely important alternative cause was found in 1.6%.

Guide		Comments
I.	***Is this a newly derived instrument* (Level IV)?**	
A.	Was validation restricted to the retrospective use of statistical techniques on the original database? (If so, this is a Level IV rule & is not ready for clinical application).	No. Validation was performed prospectively on a separate group of patients identified as requiring CT to evaluate for a possible ureteral stone.
II.	**Has the instrument been validated? (Level II or III). If so, consider the following:**	
1a	Were all important predictors included in the derivation process?	Yes. The list of factors considered for the STONE score by the authors is exhaustive, including race; presence, location, and duration of pain; presence of additional symptoms such as nausea, vomiting, and dysuria; prior personal or family history of stones; vital sign measurements; physical exam findings (specifically abdominal or back tenderness); and laboratory values.
1b	Were all important predictors present in significant proportion of the study population?	Yes. Only 3 factors were present in < 10% of the derivation population: presence of diarrhea (5.1%), family history of kidney stones (6.1%), and upper abdominal tenderness (8.8%).
1c	Does the rule make clinical sense?	Yes and no. While the components of the rule clearly reflect the risk of stones in the derivation and validation cohorts, it seems odd that race would play such a large role in the risk of renal colic (value of 3 points).
2	Did validation include prospective studies on several different populations from that used to derive it (II) or was it restricted to a single population (III)?	No. Validation was restricted to a single population at the same two centers from which the derivation cohort was obtained. The external validity of this rule has yet to be established.
3	*How well did the validation study meet the following criteria?*	
3a	Did the patients represent a wide spectrum of severity of disease?	Uncertain. The authors provide no information with regards to stone size or the presence and degree of hydronephrosis. Additionally, we are given no information with regards to the need for surgical or procedural intervention in patients in either cohort (i.e. lithotripsy, stent placement, surgical retrieval).
3b	Was there a blinded assessment of the gold standard?	Yes. CT was used as the gold standard. For the derivation cohort, "factors were then abstracted from the medical records blinded to CT reports", and "we blindly abstracted and categorized the results of the dictated CT reports." (p. 2)

<table>
<tr><td></td><td></td><td>For the validation cohort “the research associated recorded all relevant factors (listed in supplementary appendix 1) from the derivation phase for the enrolled patients before the results of the CT were known,” and “the CT result was categorized blinded to the clinical factors (except laterality of pain) and point total.” (p. 3)</td></tr>
<tr><td>3c</td><td>Was there an explicit and accurate interpretation of the predictor variables & the actual rule without knowledge of the outcome?</td><td>Yes. See above.</td></tr>
<tr><td>3d</td><td>Did the results of the assessment of the variables or of the rule influence the decision to perform the gold standard?</td><td>No. For the retrospective derivation cohort, only patients with a CT “flank pain protocol” were eligible for inclusion. For the validation cohort, only patients “in whom the clinician intended to obtain a CT scan for kidney stone” were eligible for inclusion.</td></tr>
<tr><td>4</td><td>How powerful is the rule (in terms of sensitivity & specificity; likelihood ratios; proportions with alternative outcomes; or relative risks or absolute outcome rates)?</td><td>In the derivation cohort:
• The STONE score had a misclassification rate of 0.23 (95% CI 0.22-0.23) and an AUC of 0.82 (0.74-0.90).
• The ureteral stone prevalence based on risk group is shown in Table 2.
• Acutely important alternative causes were found on CT in 2.9% of the cohort (0.3% in the high probability group).

In the validation cohort:
• The STONE score had an AUC of 0.79 (95% CI 0.76-0.83).
• The ureteral stone prevalence based on risk group is shown in Table 2.
• Acutely important alternative causes were found on CT in 3.7% of the cohort (1.6% in the high probability group).

Table 2. Prevalence of ureteral stone by STONE score category (STONE score range)
<table>
<tr><th></th><th>Derivation cohort</th><th>Validation cohort</th></tr>
<tr><td>Low (0-5)</td><td>8.3%</td><td>9.2%</td></tr>
<tr><td>Moderate (6-9)</td><td>51.6%</td><td>51.3%</td></tr>
<tr><td>High (10-13)</td><td>89.6%</td><td>88.6%</td></tr>
</table></td></tr>
<tr><td><b>III.</b></td><td><b>Has an impact analysis</b></td><td></td></tr>
</table>

	demonstrated change in clinical behavior or patient outcomes as a result of using the instrument? (Level I). If so, consider the following:	
1	How well did the study guard against bias in terms of differences at the start (concealed randomization, adjustment in analysis) or as the study proceeded (blinding, co-intervention, loss to follow-up)?	Well. The authors did a good job of blinded the research associated. This involved both blinding those abstracting the clinical factors to the CT results and blinding those abstracting CT results to the clinical factors and STONE score calculations. While a convenience sample was obtained for the validation cohort during "defined periods," the authors do note that this included "overnights, weekends, and holidays."
2	What was the impact on clinician behavior and patient-important outcomes?	No impact analysis was performed.

Limitations:

1. **The authors do not provide a flowchart of patients eligible for inclusion in the validation set. It is unclear how many eligible patients were approached, how many were enrolled, and if there were any differences between those enrolled and those not enrolled.**

2. **The STONE score was derived and validated at the same two EDs, and will need external validation prior to widespread use.**

3. **No impact analysis has been performed to determine how to clinically employ the results of the STONE score, and verify a benefit with respect to patient-centered outcomes.**

Bottom Line:

The authors in this study retrospectively derived, then prospectively validated, a clinical score using five factors most associated with the presence of a ureteral stone: male sex, acute onset of pain, non-black race, presence of nausea or vomiting, and microscopic hematuria. This rule accurately predicted the risk of ureteral stone on CT, and the risk of a clinically important alternative diagnosis was low in the high-risk group, at 1.6%. These results will need to be validated in additional settings, and the impact of the STONE score on diagnostic imaging and patient-centered outcomes will need to be assessed in order to define a role for the score.