Journal of Pediatric Urology (2017) 13, 192–198

Reliability of grading of vesicoureteral reflux and other findings on voiding cystourethrography

Anthony J. Schaeffer [a], Saul P. Greenfield [b], Anastasia Ivanova [c], Gang Cui [c], J. Michael Zerin [d], Jeanne S. Chow [e,f], Alejandro Hoberman [g], Ranjiv I. Mathews [h], Tej K. Mattoo [i], Myra A. Carpenter [c], Marva Moxey-Mims [j], Russell W. Chesney [k,†], Caleb P. Nelson [e]

Correspondence to: A.J. Schaeffer, Primary Children's Hospital, Pediatric Urology, Suite 2200, 100 Mario Capecchi Drive, Salt Lake City, UT 84132, USA

aschaeffer78@gmail.com (A.J. Schaeffer)

Keywords
Vesico-ureteral reflux; Voiding cystourethrogram; Urinary tract infection; Radiology; Concordance; Classification

Received 10 March 2016
Accepted 30 June 2016
Available online 6 September 2016

Summary

Introduction
Voiding cystourethrography (VCUG) is the modality of choice to diagnose vesicoureteral reflux (VUR). Although grading of VUR is essential for prognosis and clinical decision-making, the inter-observer reliability for grading has been shown to vary substantially. The Randomized Intervention for Children with VesicoUreteral Reflux (RIVUR) trial provides a large cohort of children with VUR to better understand the reliability of VCUG findings.

Objective
To determine the inter-observer consistency of the grade of VUR and other VCUG findings in a large cohort of children with VUR.

Study design
The RIVUR trial is a randomized controlled trial of antimicrobial prophylaxis in children with VUR diagnosed after UTI. Each enrollment VCUG was read by a local clinical (i.e. non-reference) radiologist, and independently by two blinded RIVUR reference radiologists. Reference radiologists' disagreements were adjudicated for trial purposes. The grade of VUR and other VCUG findings were extracted from the local clinical radiologist's report. The unit of analysis included individual ureters and individual participants. We compared the three interpretations for grading of VUR and other VCUG findings to determine the inter-observer reliability.

Results
Six-hundred and two non-reference radiology reports from 90 institutions were reviewed and yielded the grade of VUR for 560 left and 524 right ureters. All three radiologists agreed on VUR grade in only 59% of ureters; two of three agreed on 39% of ureters; and all three disagreed on 2% of ureters (Table). Agreement was better (≥92%) for other VCUG findings (e.g. bladder shape "normal"). The non-reference radiologists' grade of VUR differed from the reference radiologists' adjudicated grade by exactly one grade level in 19% of ureters, and by two or more grade levels in 2.2% of ureters. When the participant was the unit of analysis, all three radiologists agreed on the grade of VUR in both ureters in just 43% of cases.

Discussion
Our study shows considerable and clinically relevant variability in grading VUR by VCUG. This variability was consistent when comparing non-reference to the adjudicated reference radiologists' assessment and the reference radiologists to each other. This study was limited to children with a history of UTI and grade I–IV VUR and may not be generalizable to all children who have a VCUG.

Conclusion
The considerable inter-observer variability in VUR grading has both research and clinical implications, as study design, risk stratification, and clinical decision-making rely heavily on grades of VUR.

[a] Department of Surgery, Division of Pediatric Urology, University of Utah School of Medicine, Salt Lake City, UT, USA [b] Department of Pediatric Urology, Women & Children's Hospital of Buffalo, Buffalo, NY, USA [c] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA [d] Department of Radiology, Children's Hospital of Michigan, Detroit, MI, USA [e] Department of Urology, Boston Children's Hospital, Boston, MA, USA [f] Department of Radiology, Boston Children's Hospital, Boston, MA, USA [g] Department of Pediatrics, Children's Hospital of Pittsburgh, Pittsburgh, PA, USA [h] Department of Urology, Johns Hopkins Hospital, Baltimore, MD, USA [i] - Department of Pediatrics, Children's Hospital of Michigan, Detroit, MI, USA [j] National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA [k] Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN, USA † Deceased author.

http://dx.doi.org/10.1016/j.jpurol.2016.06.020

Table Study summary

Characteristic		
No. of VCUG reports analyzed	602	
Gender of participants		
Male	49	
Female	553	
Age in months at time of VCUG (median) [IQR]	11	[5,30]
No. of ureters analyzed	1081	
Reflux grade agreement		
Between non-reference and each reference radiologist (three-way)		
All three agree	638/1081	(59%)
Two agree, one disagree	417/1081	(39%)
All three disagree	27	(2%)
Between non-reference and adjudicated reference radiologists' score (two-way)		
Agree	805	(75%)
Disagree	275	(25%)
Kappa (95% CI)	0.66	(0.62—0.69)

Introduction

Voiding cystourethrography (VCUG) is the modality of choice to diagnose vesicoureteral reflux (VUR) and certain other urinary tract abnormalities in children. VCUG permits grading of VUR using the five-level International Reflux Scale (IRS) [1], and grade of VUR is strongly associated with outcomes such as spontaneous resolution, recurrence of urinary tract infection (UTI), renal scarring, and others [2—4]. Recent guidelines recommend clinical decision-making based on the grade of VUR on VCUG, including observation (without medical therapy) for selected children with grade I or II VUR [5].

While both research findings and clinical recommendations assume that the IRS grading scale is reliable and reproducible, this may not be the case, as grade discrepancies between readers occur in significant numbers of children with VUR [6—8]. However, there are limitations to the published data, and more rigorous measurement of inter-observer agreement regarding VCUG findings is needed.

The Randomized Intervention for Children with VesicoUreteral Reflux (RIVUR) trial is a randomized double-blind placebo-controlled trial of antimicrobial prophylaxis among children with grade I—IV VUR and urinary tract infection [9]. Each child included in RIVUR had a VCUG at study entry, and each VCUG was independently reviewed by a local clinical (non-reference) radiologist, as well as by two RIVUR reference radiologists.

The purpose of this study was to assess inter-observer agreement for grade of VUR and other VCUG findings among RIVUR participants.

Materials and methods

The RIVUR trial randomized 607 children with VUR to trimethoprim-sulfamethoxazole prophylaxis or placebo [9]. Three assessments of pre-enrollment VCUG images were performed: one by the non-reference radiologist who performed the initial clinical interpretation, and one by each of the two reference pediatric radiologists designated to read VCUGs for the RIVUR trial (JMZ, JSC). The RIVUR reference radiologists assessed grade of VUR and other findings including but not limited to bladder (i.e. shape, trabeculation) and urethral anatomy. Disagreements between reference radiologists were adjudicated. Reference radiologists accessed original digital images placed on compact discs in the DICOM format and viewed them on high-resolution monitors. There was no standardized reader software.

The non-reference radiology reports were reviewed for findings besides VUR using the same data form used by the reference radiologists. Two differences exist between the non-reference and reference assessments. First, although each reference radiologist was required to provide an assessment for every item on the RIVUR data report form, the local (non-reference) radiologists were under no such requirements. Accordingly, not every item assessed by the RIVUR reference radiologists was included in every non-reference VCUG report. If no statement was made about a certain item (e.g. left but not the right ureter grade of VUR) in the non-reference report, the field for that item was coded as missing. However, all other items mentioned in the report were analyzed. Second, the non-reference radiologist could report intermediate grades of VUR (e.g. grade II—III), whereas the reference radiologists were only permitted to use grades I, II, III, IV, or V. Of ureters graded by non-reference radiologists, 8% (90/1080) received such an intermediate grade. To handle this discrepancy analytically, intermediate grades of reflux assigned by non-reference radiologists were given 0.5 units for each grade they encompassed. For example, 0.5 units were assigned to grade II and 0.5 units to grade III for non-reference grade II—III VUR.

A three-way comparison between the non-reference radiologist and each reference radiologist was used to determine the number and percent of responses in which all three radiologists agreed, two radiologists agreed and one disagreed, or all three radiologists disagreed. Similarly,