Journal of Pediatric Urology (2014) 10, 107–111

Journal of Pediatric Urology

Reliability of voiding cystourethrogram for the grading of vesicoureteral reflux

Brock B. O'Neil [a,*], Patrick C. Cartwright [a], Constance Maves [b], Karin Hoeg [b], Angela P. Presson [c], M. Chad Wallis [a]

[a] *Division of Urology, University of Utah, Primary Children's Medical Center, 30 North 1900 East, Salt Lake City, UT 84132, USA*
[b] *Department of Medical Imaging, Primary Children's Medical Center, 100 N Mario Capecchi Dr, Salt Lake City, UT 84113, USA*
[c] *Division of Epidemiology, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84108, USA*

Received 8 March 2013; accepted 28 June 2013
Available online 25 July 2013

KEYWORDS
Vesicoureteral reflux;
Reproducibility of results;
Classification;
Radiology

Abstract *Objective:* The voiding cystourethrogram (VCUG) is a commonly employed radiographic test used in the management of vesicoureteral reflux (VUR). Recently, the reliability of VCUG to accurately grade VUR has been questioned. The purpose of this study is to examine reliability of the VCUG for the grading of VUR in a setting mimicking daily practice in a busy pediatric hospital.
Materials and methods: Two-hundred consecutive VCUGs were independently graded by two pediatric urologists and two pediatric radiologists according to the International Classification of Vesicoureteral Reflux. A weighted kappa coefficient was calculated to determine inter-rater agreement and a modified McNemar test was performed to assess rater bias. Further assessment for impact on clinical and research decision-making was made for disagreement between grades II and III.
Results: Weighted kappa values reflect strong reliability of VCUG for grading VUR between and among urologists and radiologists ranging from 0.95 to 0.97. There was statistically significant bias with radiologists reporting higher grades. Despite high kappa values, disagreement between raters was not infrequent and most common for grades II–IV.
Conclusions: VCUG is reliable for grading VUR, but small differences in grading between raters were detected and may play an important role in clinical decision-making and research outcomes.

* Corresponding author. Tel.: +1 801 213 2700.
E-mail address: brock.oneil@hsc.utah.edu (B.B. O'Neil).

1477-5131/$36
http://dx.doi.org/10.1016/j.jpurol.2013.06.014

Introduction

The voiding cystourethrogram (VCUG) is commonly used in the evaluation of urinary tract infections and in the setting of hydronephrosis to detect vesicoureteral reflux (VUR). The International Classification of Vesicoureteral Reflux (ICVUR) has been the most widely used grading system. Previous work evaluating the use of VCUG for the detection of VUR has focused on timing of examination [1,2], need for repeat filling and voiding cycles [3], and reliability of digital interpretation equipment [4].

One study from more than a decade ago examined inter-rater reliability between three radiologists and found high agreement as assessed by kappa values [5]. However, the reliability of VCUG for accurate diagnosis of VUR grade has been called into question with disagreement between raters in early stages of the Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) trial [6]. Metcalfe et al. then formally assessed inter-rater reliability and reported poor to modest agreement for pediatric urologists, radiologists and senior urology residents [7]. Each of the 13 raters graded 28 VCUGs in an artificial research setting with an average inter-rater reliability of only 0.53.

In addition to understanding the reliability of VCUG in order to appropriately interpret research results, it is also important to understand differences to provide best clinical care. For example, pediatricians depend heavily on radiologists' reports for management and referral decisions whereas pediatric urologists tend to base treatment on their own interpretation. The purpose of this study is to examine reliability of the VCUG for the grading of VUR in a setting mimicking daily practice in a busy pediatric hospital. We hypothesized that the ICVUR system for grading VUR has acceptable inter-rater reliability in this situation.

Materials and methods

Two-hundred consecutive VCUGs were independently graded by two fellowship-trained pediatric urologists and two fellowship-trained pediatric radiologists after approval by our Institutional Review Board. Raters were provided with a list of two unique de-identified numbers to query images from our hospital imaging archive in a manner blinded to the patients' clinical history. VUR was graded from 0 to V according to the ICVUR. Raters briefly reviewed this grading system together for 10 min prior to initiation of the study review period. VCUGs were graded in an environment replicating daily imaging review practice. That is, pediatric radiologists reviewed studies on dedicated radiology PACS imaging stations and the pediatric urologists reviewed studies on a standard computer monitor in a clinic setting using the hospital's archival and imaging software.

As a result of the higher expected incidence of negative VCUGs and to provide adequate variability, only one out of four VCUGs with no VUR in either unit based on grading of the original radiologist was included. VCUGs were also excluded if ordered primarily for neurogenic bladder, exstrophy, or posterior urethral valve. Patients with other anatomic abnormalities including ectopic ureterocele and ureteral duplication were not excluded. In total, 394 VCUGs performed at our institution between November 1, 2011 and March 1, 2012 were screened. Studies were excluded for above indications and one out of four VCUGs with no reflux in either unit were included resulting in a total of 200 studies. These 200 VCUGs were then evaluated independently by the four raters at their own pace over a six week period.

Inter-rater reliability among grades 0—V and grades I—V was assessed between a) the two pediatric radiologists, b) the two pediatric urologists, c) the radiologists and urologists, d) the averaged urologists' scores and the original radiologist rater, and e) the averaged radiologists' scores and the original radiologist rater; for a total of ten evaluations. For each comparison, a quadratic weighted kappa coefficient was calculated to determine inter-rater agreement, where values between 0.81 and 1 were considered near perfect agreement [8]. The median quadratic weighted kappa coefficient and its 95% CI were estimated from 2000 bootstrap replicates, where the sampling unit was the VCUG in order to account for the correlation between renal images within a VCUG (as renal units within a patient were likely more correlated on average because of the degree of reflux and/or the image quality). Note that to compute the averaged scores, we followed averaging by rounding in *R* using the `round()` function, which implements the IEC 60559 standard: '*go to the even digit*' to prevent the rounded scores from exhibiting an upward bias.

Intra-rater agreement is generally higher than inter-rater agreement, and some authors argue it is unnecessary to assess intra-rater agreement if inter-rater agreement is high [9]. Study design called for assessment of intra-rater agreement only if inter-rater agreement fell below 0.81.

For the grades I—V analysis, VCUGs were selected that had VUR >0 according to the original rater. A modified McNemar test was performed to assess for systematic rater bias where values near or at 0.5 indicated low bias. Finally, we evaluated agreement at the cutoff between grades II and III as this has been suggested as an important clinical cutoff [10—12].

All statistical analyses were performed in R v. 2.15.0 (http://www.R-project.org/); tests were two-tailed and an alpha level of 0.05 was used to assess significance.

Results

Seventy-five percent of VCUGs were in female children with a median age of 47.9 months (range 0.3—182.7). Original interpretation was performed by one of 14 pediatric radiologists at our institution with a median of 17 (range 8—33) images available per study. Table 1 shows that almost half of renal units evaluated did not have any VUR according to the original rater. The remaining units had relatively equal proportions of grades I—III with grades IV and V being less common. Fig. 1 shows that the urologists tended to grade studies more commonly as II and radiologists were more likely to score studies as III—IV. This is further substantiated by a significant McNemar test, indicating bias towards higher grading by radiologists compared to urologists (Table 2).

Weighted kappa values reflect strong reliability of VCUG for grading VUR between and among urologists and radiologists (Table 2). When excluding renal units scored as a 0 by the original interpreting radiologist, weighted kappa values remained high (0.93, 95% CI 0.91—0.95). Both the urologists and radiologists exhibited high agreement with the original

Table 1 Baseline study characteristics based on the original radiologists' interpretation.

	Left	Right
Total renal units	200	200
Non-standard[a]	6	6
Grade 0	111	88
Grade I	21	24
Grade II	23	30
Grade III	24	33
Grade IV	11	14
Grade V	4	5

[a] Radiologist used a non-standard grading score.

interpreting radiologist with kappa values of 0.93 for both comparisons. As inter-rater agreement was found to be >0.81, intra-rater agreement was not assessed.

Visual representation of agreement between rater scores is shown in Fig. 2. Closer inspection of scoring reveals important discrepancies in grading clustered around grades II–IV. For example, Plot A represents agreement between the two pediatric urologists. Urologist #1 rated 50 renal units as grade III (13 + 30 + 6 + 1) and Urologist #2 scored 39 renal units as grade III (3 + 30 + 6). They agreed with each other on 30 of the same renal units (highlighted in green). This indicates that in up to 49% of ratings assigned as III by one of the two urologists, the other disagreed.

As differential treatment is suggested by some for grades II or less compared with III or greater [13], a more clinically relevant distinction may be how often a VCUG is under-graded by the other reviewer. That is, how often did one rater score a unit as III but the other rater assign a lower score. For the urologists, 16 units were graded as II by one rater out of 46 (35%) units graded as III and 15 out of 51 (29%) for the radiologists. Possible over-grading, or how often did one rater score a unit as II but the other assign a higher score, is also pertinent. For the urologists, this is 16/71 (23%) and for the radiologists, 15/40 (38%).

When disagreement did occur, the degree of difference between grading was generally by a difference of one (yellow highlighting, Fig. 2). However, radiologists were more likely to differ by more than one grade (9 units) compared with urologists (1 unit, OR 0.11, $p = 0.02$, red highlighting).

Discussion

Although no study evaluating reliability can completely replicate daily practice, evaluation of VCUG in this study that attempted to reproduce daily clinical practice as much as possible shows it to be very reliable for the overall grading of VUR based up high kappa values. This was previously reported by some authors [5] but not confirmed in the work of others [7]. Metcalfe et al. showed high level agreement for grade I (0.98) and acceptable agreement for grade V (0.72) but poor agreement for the other grades. Further, agreement when assessed as an aggregate for all grades ranged between 0.56 and 0.61 [7].

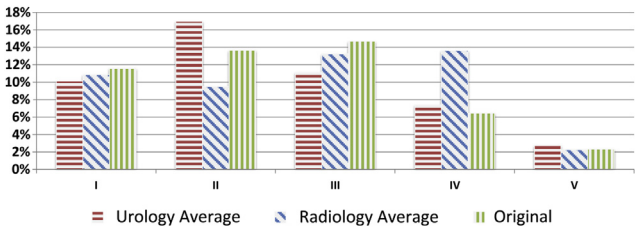

Figure 1 Proportion assigned to each grade of VUR by interpreting group.

There are clear methodological differences between this work and the study by Metcalfe et al. that likely explains the differences in results. First, raters in the present study briefly reviewed the ICVUR schema prior to beginning the study in a group setting, whereas, in earlier work, raters reviewed a single PowerPoint slide depicting and explaining the grading system. It is possible that this simple intervention might increase rater reliability and regular collective review of this schema may help maintain reliability for research protocols that depend highly on reliable VCUG grading. Next, the sheer volume of VCUGs reviewed in this work that evaluated 200 studies per rater compared with 28 in the previous work may have an impact on rater reliability. Finally, we sought to replicate daily practice for reviewing VCUGs. Metcalfe et al. displayed images copied into PowerPoint, which has the potential to limit display resolution, operator manipulation of images, and the number of images available for review. This work utilized the hospital's imaging software and archival system with standard imaging display equipment, allowing each interpreter to review the entire series of images captured during the VCUG procedure and manipulate the images as needed to arrive at the appropriate grade.

Despite high kappa values, disagreement between raters in this study was not an uncommon event with differences being generally by only a single grade (see Fig. 2). However, scoring discrepancies were more common with certain grades (II–IV). It is in this critical range of grades that variation in grading may result in important alterations to clinical decision-making and to interpretation of research results that rely on the VCUG grade for patient stratification.

One potential explanation for the high kappa values but with some disagreement in grades II–IV is that half of the renal units assessed did not have VUR, and agreement among all raters on grade 0 VUR is extremely high. Yet, when we repeated the weighted kappa measures excluding studies where the original radiologist scored the renal unit as 0, kappa remained high (see Table 2).

Understanding the nature of the quadratic weighted kappa coefficient compared with absolute agreement is also important in understanding this apparent discrepancy. The quadratic weighted kappa coefficient is more strongly negatively affected by large discrepancies in grading than by small ones. That is, kappa is much lower (closer to 0) when one rater assigns grade I and another assigns grade V compared with the situation when one assigns I and the other II. In this study, wide discrepancies that would bring the kappa down were very uncommon. Differences in grading were not uncommon, but differences were generally small. As a result, these instances of disagreement did not have a large impact on kappa.

Recently, some have argued to initially limit treatment and follow-up imaging for those diagnosed with grades I or II VUR [13]. If this is followed, differences in diagnosis at this cutoff are critical for determining treatment. When treatment

Table 2 Reliability and Bias scores by rater group.

	Kappa VUR grades 0–V		Kappa VUR grades I–V		McNemar		
	Value	95% CI	Value	95% CI	Ratio	χ^2	p
Urologist 1 vs. urologist 2	0.97	0.96–0.98	0.96	0.94–0.97	0.53	0.2	0.68
Radiologist 1 vs. radiologist 2	0.95	0.93–0.97	0.94	0.91–0.96	0.59	1.7	0.19
Avg urologists vs. avg radiologists[a]	0.95	0.93–0.96	0.93	0.91–0.95	0.88	42.4	<0.0001
Urologists vs. original rater[b]	0.95	0.94–0.96	0.93	0.91–0.95	0.65	6.1	0.01
Radiologists vs. original rater[c]	0.95	0.94–0.96	0.93	0.92–0.95	0.27	17.3	<0.0001

Comparing the rounded average rating of.
[a] Urologists to radiologists.
[b] Urologists to the original radiologist rater.
[c] Radiologists to the original radiologist rater.

A

Urologist #1 \ Urologist #2	0	I	II	III	IV	V
0	203	3	0	0	0	0
I	4	32	8	0	0	0
II	0	2	55	3	0	0
III	0	0	13	30	6	1
IV	0	0	0	6	20	7
V	0	0	0	0	0	7

B

Radiologist #1 \ Radiologist #2	0	I	II	III	IV	V
0	196	4	4	0	0	0
I	1	37	1	0	3	0
II	2	4	25	4	0	0
III	0	0	11	36	12	0
IV	0	0	0	6	41	6
V	0	0	0	0	0	6

C

Urologists \ Radiologists	0	I	II	III	IV	V
0	199	9	2	0	0	0
I	1	29	2	0	0	0
II	1	5	43	27	5	0
III	0	0	0	10	20	0
IV	0	0	1	0	39	0
V	0	0	0	0	1	6

D

Urologists \ Original Radiologist Rater	0	I	II	III	IV	V
0	195	11	3	0	0	0
I	4	27	1	0	0	0
II	1	7	48	23	0	0
III	0	0	1	24	3	0
IV	0	0	0	10	22	2
V	0	0	0	0	0	6

E

Radiologists \ Original Radiologist Rater	0	I	II	III	IV	V
0	190	10	1	0	0	0
I	9	31	2	0	0	0
II	1	4	37	4	1	0
III	0	0	13	22	0	0
IV	0	0	0	31	24	3
V	0	0	0	0	0	5

Figure 2 Comparison of reflux grading between A. Urologist 1 and Urologist 2, B. Radiologist 1 and Radiologist 2, C. the average of urologists 1 and 2 (Urologists) with the average of radiologists 1 and 2 (Radiologists), D. the averaged urologists' scores and the original radiologist rater, and E. the averaged radiologists' scores and the original radiologist rater. Green highlighting indicates perfect agreement, yellow indicates a reflux grade difference between raters of one and red indicates a reflux grade difference of two or more between raters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is based on the grading of a single rater, under-grading and possible under-treatment would occur in 23%–38% and over-grading and possible over-treatment might occur in 29%–35%. If it is felt that this cutoff is truly of clinical importance, one possible strategy to reduce inappropriately assigning treatment would be to have a second rater review any VCUG scored as grade II or grade III, and agree together which is the appropriate grade. This is similar to the process done to assign grades at entry in the RIVUR trial [7], but limits review to those studies that are at highest risk for having clinically meaningful disagreement rather than having independent review of every VCUG.

Finally, we found statistically significant bias with both the study and the original radiologists tending to grade VUR higher than the urologists. As primary care providers likely depend on the radiologists' written report, we speculate that this higher grading tendency may have the added benefit of a more prompt referral to a specialist.

Conclusion

This work confirmed that VCUG is reliable for grading VUR when evaluated in a setting parallel to clinical practice. However, disagreement when grading VUR based on VCUG by experienced raters does occur and has the potential to have meaningful impact on treatment outcomes. Combined review of images by multiple reviewers as occurs for study entry in the RIVUR trial has the potential to reduce this impact and may play a role in both research protocols and clinical practice.

Funding

This investigation was supported by the University of Utah Study Design and Biostatistics Center, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 8UL1TR000105 (formerly UL1RR025764).

Conflict of interest

None.

References

[1] Sathapornwajana P, Dissaneewate P, McNeil E, Vachvanichsanong P. Timing of voiding cystourethrogram after urinary tract infection. Arch Dis Child 2008;93:229–31.

[2] Thompson M, Simon SD, Sharma V, Alon US. Timing of follow-up voiding cystourethrogram in children with primary vesicoureteral reflux: development and application of a clinical algorithm. Pediatrics 2005;115:426—34.
[3] Jequier S, Jequier JC. Reliability of voiding cystourethrography to detect reflux. Am J Roentgenol 1989;153:807—10.
[4] Kronemer KA, Don S, Luker GD, Hildebolt C. Soft-copy versus hard-copy interpretation of voiding cystourethrography in neonates, infants, and children. Am J Roentgenol 1999;172:791—3.
[5] Craig JC, Irwig LM, Christie J, Lam A, Onikul E, Knight JF, et al. Variation in the diagnosis of vesicoureteric reflux using micturating cystourethrography. Pediatr Nephrol 1997;11:455—9.
[6] Greenfield SP, Carpenter MA, Chesney RW, Zerin JM, Chow J. The RIVUR voiding cystourethrogram pilot study: experience with radiologic reading concordance. J Urol 2012;188:1608—12.
[7] Metcalfe CB, MacNeily AE, Afshar K. Reliability assessment of international grading system for vesicoureteral reflux. J Urol 2012;188:1490—2.
[8] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159—74.
[9] Karanicolas PJ, Bhandari M, Kreder H, Moroni A, Richardson M, Walter SD, et al. Evaluating agreement: conducting a reliability study. J Bone Jt Surg Am 2009;91(Suppl. 3):99—106.
[10] Montini G, Rigon L, Zucchetta P, Fregonese F, Toffolo A, Gobber D, et al. Prophylaxis after first febrile urinary tract infection in children? A multicenter, randomized, controlled, noninferiority trial. Pediatrics 2008;122:1064—71.
[11] Garin EH, Olavarria F, Garcia Nieto V, Valenciano B, Campos A, Young L. Clinical significance of primary vesicoureteral reflux and urinary antibiotic prophylaxis after acute pyelonephritis: a multicenter, randomized, controlled study. Pediatrics 2006; 117:626—32.
[12] Roussey-Kesler G, Gadjos V, Idres N, Horen B, Ichay L, Leclair MD, et al. Antibiotic prophylaxis for the prevention of recurrent urinary tract infection in children with low grade vesicoureteral reflux: results from a prospective randomized study. J Urol 2008;179:674—9 discussion 9.
[13] Peters CA, Skoog SJ, Arant Jr BS, Copp HL, Elder JS, Hudson RG, et al. Summary of the AUA guideline on management of primary vesicoureteral reflux in children. J Urol 2010;184:1134—44.