Quantitative Ultrasound for Measuring Obstructive Severity in Children with Hydronephrosis

Juan J. Cerrolaza,* Craig A. Peters,* Aaron D. Martin, Emmarie Myers, Nabile Safdar and Marius George Linguraru

From Sheikh Zayed Institute for Pediatric Surgical Innovation (JJC, CAP, ADM, EM, MGL), and Division of Urology and Pediatrics (CAP, ADM), Children’s National Health System, and Departments of Radiology and Pediatrics, School of Medicine and Health Science, George Washington University (MGL), Washington, D.C., and Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, Georgia (NS)

Purpose: We define sonographic biomarkers for hydronephrotic renal units that can predict the necessity of diuretic nuclear renography.

Materials and Methods: We selected a cohort of 50 consecutive patients with hydronephrosis of varying severity in whom 2-dimensional sonography and diuretic mercaptoacetyltriglycine renography had been performed. A total of 131 morphological parameters were computed using quantitative image analysis algorithms. Machine learning techniques were then applied to identify ultrasound based safety thresholds that agreed with the t₁/₂ for washout. A best fit model was then derived for each threshold level of t₁/₂ that would be clinically relevant at 20, 30 and 40 minutes. Receiver operating characteristic curve analysis was performed. Sensitivity, specificity and area under the receiver operating characteristic curve were determined. Improvement obtained by the quantitative imaging method compared to the Society for Fetal Urology grading system and the hydronephrosis index was statistically verified.

Results: For the 3 thresholds considered and at 100% sensitivity the specificities of the quantitative imaging method were 94%, 70% and 74%, respectively. Corresponding area under the receiver operating characteristic curve values were 0.98, 0.94 and 0.94, respectively. Improvement obtained by the quantitative imaging method over the Society for Fetal Urology grade and hydronephrosis index was statistically significant (p<0.05 in all cases).

Conclusions: Quantitative imaging analysis of renal sonograms in children with hydronephrosis can identify thresholds of clinically significant washout times with 100% sensitivity to decrease the number of diuretic renograms in up to 62% of children.

Key Words: biological markers, hydronephrosis, kidney, technetium Tc 99m mertiatide, ultrasonography

Ultrasound is the mainstay of imaging for pediatric hydronephrosis yet is limited by its subjectivity and lack of apparent correlation with functional imaging modalities. To improve the clinical usefulness of sonography for hydronephrosis, we developed a semiautomatic quantitative imaging analysis algorithm to objectively describe hydronephrosis and to define ultrasound based thresholds of obstruction below which diuretic renography can be safely avoided. Quantitative imaging

Abbreviations and Acronyms

D = dimensional
DR = diuretic renography
HI = hydronephrosis index
LOG = logistic regression analysis
MAG3 = mercaptoacetyltriglycine
QI = quantitative imaging
SFU = Society for Fetal Urology
SVM = support vector machine
US = ultrasound
techniques rely on analysis of digital images and creation of normative models of shape and appearance to which a test case is compared. The test case can then be defined, based on probability, as being within a specific range of values different from the norm. By using machine learning to correlate patterns of hydronephrosis with the diuretic renography parameters, in this case $t_{1/2}$ washout, the correlation between structure and function can be defined and used to categorize subsequent test cases. With the primary goal being to identify ultrasound derived thresholds of safety for hydronephrotic renal units where diuretic renography had been performed. The quantitative imaging algorithm was then assessed as a predictor of the washout parameters of diuretic renography.

Currently the most widely used instrument for grading hydronephrosis sonographically is the SFU system, although it is not universally accepted as useful. Shapiro and Venkatesan et al proposed a quantitative measure of hydronephrosis severity, the HI, defined as the ratio of the collecting system area to the total area of the collecting system and renal parenchyma. However, the practical use of this index remains unclear. While different clinicians use differing thresholds for selecting patients in whom to perform DR and for defining clinically significant obstruction, it would seem valuable to perform DR only in those in whom it is likely to reveal clinically relevant parameters for intervention or close followup.

We hypothesized that hydronephrosis can be objectively quantified sonographically using modern imaging analysis, and that there is a correlation between degree of hydronephrosis and severity of obstruction as measured by DR. This correlation likely is not linear, but demonstrates a threshold of dilatation above which clinically significant drainage abnormalities are detectable on DR. If this relationship can be defined, then quantitative assessment of degree of hydronephrosis on US could be used to selectively identify which patients have more severe obstruction (by DR parameters) and should undergo DR, and which patients have less severe hydronephrosis and should forgo DR.

**MATERIALS AND METHODS**

**Study Population**

Under an internal review board approved protocol we retrospectively selected 50 consecutive patients (32 males and 18 females) with hydronephrosis who had undergone concurrent renal 2-D ultrasound and DR (MAG3 scan). Mean patient age was 9.6 months (median 4.0, range 0 to 168). Of the cases 21 were left sided, 29 were right sided and 8 were bilateral. Cases involving vesicoureteral reflux, ureterovesical junction obstruction, ureterocele or hydroureter/megaureter were excluded to avoid confounding factors. Ultrasound acquisition was performed using the Philips® iU22 (C2-9-D transducer) and GE® Logiq E9 (C8-5 transducer) scanners between July 2012 and January 2013. Images were identified that contained the closest to a full coronal view that included the renal pelvis and calyces. MAG3 scan was performed using a standardized method described previously. All patients underwent DR based on clinician preference, and no study was postoperative.

**Quantitative Imaging Analysis**

Manual segmentation of the renal parenchyma and collecting system was performed to create a cohort for program calibration and algorithm development. This was accomplished by manually drawing the best line around the perimeter of the renal parenchyma and the collecting system (using the ITK-Snap Medical Image Segmentation Tool, http://www.itk-snap.org), excluding the extrarenal pelvis. The intrarenal pelvis was defined as the area falling within a line drawn around the hilar curve of the parenchyma (fig. 1). All segmentations were reconciled with 2 observers (pediatric radiologist and urologist) and submitted for QI analysis. The same cohort underwent grading using the SFU classification by a pediatric radiologist and a pediatric urologist.

A total of 131 morphological parameters were computed (fig. 2 and supplementary table, http://jurology.com/). These descriptors can be intuitively grouped into 3 categories, ie size, geometric shape and curvature. Size descriptor parameters describe the size of the collecting system and renal parenchyma, such as relative area (eg HI) and perimeter, depth of calyces, and thickness of parenchyma. Geometric shape descriptor parameters include such factors as circularity ratio, eccentricity of the circumscribed ellipse and convexity. Curvature descriptors are parameters that describe the local curvature of the structures under analysis (ie kidney and collecting system).

These quantitative shape features were used as potential predictive variables to evaluate the diagnostic performance of 2 different supervised classification models, the LOG and SVM with nonlinear kernel. Both approaches have been widely used in the field of computer-aided diagnosis with satisfactory results.

These 2 classification approaches were applied to the test cohort to compute US based specific thresholds to identify patients with a $t_{1/2}$ above different thresholds, ie 20, 30 and 40 minutes, since not all clinicians agree as to the clinically relevant DR thresholds. These cutoffs represent commonly used clinical thresholds. The design parameters of the model were adjusted to maximize the sensitivity of detecting severe cases of hydronephrosis, that is no case with $t_{1/2}$ above the threshold was misclassified. A best fit model was then derived for each threshold using optimal morphological parameters.
and SVM or LOG to categorize the renal units and severity of obstruction.

To compare this QI method to the current SFU grading system in selecting cases where DR was to be performed, we compared SFU grade 3 and above or only SFU grade 4 as the criterion for performing DR. Sensitivity and specificity were determined, as well as ROC curve analysis with AUC calculation. A similar assessment was made using the previously described HI. The proposed framework was implemented in MATLAB®, using a 64-bit 2.8 GHz processor, with an average execution time of 12 seconds per case.

Statistical Analysis

The cohort data were used to assess the performance of the QI algorithm described previously, using leave-one-out cross-validation methodology. From the 131 morphological features computed for each case an optimal subset of predictive variables was selected for each $t_1/2$ threshold using the feature selection technique proposed by Cai et al. The significance of the improvement obtained by the QI method over the SFU grade and HI was statistically verified using the McNemar test. Additionally the interobserver variance for SFU grading was assessed using the Fleiss kappa statistic.

In the absence of previous studies that related to this topic to provide an estimation of the expected effect we performed a retrospective power analysis using a 2-tailed t-test over the output probability provided by the classifiers, SVM and LOG (significance level $= 0.05$). Effect sizes were calculated using the Hedges formulation for means differences.

RESULTS

The retrospective power analysis of the cohort of 50 patients provided a power value of greater than 0.99 in all instances. SFU grades ranged from 1 to 4 (median 3). Median differential uptake on MAG3 scan was 48% (range 14% to 100%). Median $t_1/2$ was 17 minutes (range 3 to greater than 120). Differential uptake was less than 40% in 7 patients and less than 30% in 3. All patients with differential uptake less than 30% would have been identified with a $t_1/2$ threshold of 30 minutes or greater. Two patients with differential uptake less than 40% and greater than 30% had washout times of less than 20 minutes (table 1), suggesting a nonobstructive cause of the reduced uptake.

The best performance of the QI method was obtained by means of the SVM classification system for thresholds of 20 and 30 minutes, while LOG provided better performance for a threshold of 40 minutes. Because the design parameters of the model were defined to maximize the sensitivity of detecting cases of hydronephrosis over a specified

Figure 1. 2-D ultrasound slice selection and segmentation. A, selected slice containing whole longitudinal section of kidney. B, consensus manual delineation of renal parenchyma and collecting system.

Figure 2. Morphological descriptors of renal unit. Sample illustrates morphological descriptors considered as potential predictive features of hydronephrosis severity. $ab$, eccentricity of circumscribed ellipse. $P_{max}$, maximum parenchymal thickness. $P_{min}$, minimum parenchymal thickness. $PKMA_{max}$, kidney medial axis and point of maximum curvature. $PKMA_{left}$, left axis of kidney. $PKMA_{right}$, right axis of kidney. Arrows indicate curvature measures of collecting system and kidney surface. Relative area of collecting system and parenchyma is shown.
that obtained using SFU grade when an SFU grade of 3 or 4 was used to refer patients for MAG3 imaging. For the 3 different t1/2 thresholds under study the QI algorithm maximizes the reduction in MAG3 scan within the total cohort, while no critical case with severe obstruction was missed.

**DISCUSSION**

Using the QI algorithm with machine learning in pediatric patients with hydronephrosis, US can be used to predict which patients will have clinically relevant increased washout times on DR. Applying this technique to children with hydronephrosis, particularly those with prenatally diagnosed disease, could safely decrease the need for invasive and ionizing DR. Clinicians would be able to select thresholds of safety that they find acceptable.

These results have several potentially useful implications. These observations demonstrate that renal structure and function, specifically drainage, are related. Being able to use ultrasound images to infer functional data would offer decreased invasiveness, complexity and cost, with no reduction in safety and acceptable reduction in accuracy.

This system appears to be a useful clinical tool to determine which patients would benefit from DR over using the SFU scoring system alone. While this approach does not eliminate the need for DR, it permits more specific application of a costly, complex, semi-invasive imaging test with radiation exposure. Application of this modality would be most useful in infants with prenatally detected hydronephrosis, in whom there remains controversy regarding the need for imaging beyond the postnatal US. There are currently no clearly defined guidelines based on objective metrics or clinical outcomes that aid in the selection of patients for DR. Of course, interpretation of the DR remains controversial, as is the decision regarding the need and timing for surgical intervention. This study

---

**Table 1. Cases by category**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>No. Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T1/2 threshold:</strong></td>
<td></td>
</tr>
<tr>
<td>20 Mins or greater</td>
<td>19</td>
</tr>
<tr>
<td>30 Mins or greater</td>
<td>14</td>
</tr>
<tr>
<td>40 Mins or greater</td>
<td>8</td>
</tr>
<tr>
<td><strong>SFU grade:</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
</tbody>
</table>

DR parameter, no case with t1/2 above the threshold was misclassified (100% sensitivity).

For the 3 thresholds considered, ie 20, 30 and 40 minutes, and at 100% sensitivity the specificities of the method were 94%, 70% and 74%, respectively. Compared to using the SFU grading system to select patients for MAG3 scan (table 2) or the HI, this approach provides greater specificity while maintaining maximum sensitivity.

Figure 3 illustrates the ROC curves for the quantitative imaging, SFU and hydronephrosis index techniques for selecting patients with t1/2 greater than 30 minutes. Using either hydronephrosis index or SFU grading system showed an AUC of 0.78, which was significantly less robust than the 0.94 AUC using the SVM based quantitative imaging method (p < 0.001 and p = 0.04, respectively). Similar results were obtained for washout threshold values of 20 and 40 minutes. Quantitative imaging performed significantly better than hydronephrosis index (p < 0.001 for either threshold) and significantly better than the SFU grading system (p < 0.001 for t1/2 of 20 minutes and p = 0.05 for t1/2 of 40 minutes). The interobserver reliability for SFU grading between a pediatric radiologist and a pediatric urologist revealed a Fleiss kappa statistic of 0.378, indicating fair agreement.

Table 3 outlines the potential reduction in DR parameter, no case with t1/2 above the threshold was misclassified (100% sensitivity).

The system appears to be a useful clinical tool to determine which patients would benefit from DR over using the SFU scoring system alone. While this approach does not eliminate the need for DR, it permits more specific application of a costly, complex, semi-invasive imaging test with radiation exposure. Application of this modality would be most useful in infants with prenatally detected hydronephrosis, in whom there remains controversy regarding the need for imaging beyond the postnatal US. There are currently no clearly defined guidelines based on objective metrics or clinical outcomes that aid in the selection of patients for DR. Of course, interpretation of the DR remains controversial, as is the decision regarding the need and timing for surgical intervention. This study

---

**Table 2. QI vs SFU grade or HI to identify washout time thresholds in patients with hydronephrosis**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Pos Predictive Value (95% CI)</th>
<th>Accuracy (95% CI)</th>
<th>AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T1/2 greater than 20 mins:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QI (SVM)</td>
<td>100 (82–100)</td>
<td>94 (79–99)</td>
<td>90 (70–99)</td>
<td>96 (86–99)</td>
<td>98 (92–100)</td>
</tr>
<tr>
<td>SFU grade 2</td>
<td>100 (82–100)</td>
<td>0 (0–11)</td>
<td>38 (25–53)</td>
<td>38 (25–53)</td>
<td>74 (57–83)</td>
</tr>
<tr>
<td>HI</td>
<td>100 (82–100)</td>
<td>52 (33–70)</td>
<td>56 (38–73)</td>
<td>70 (55–82)</td>
<td>77 (58–86)</td>
</tr>
<tr>
<td><strong>T1/2 greater than 30 mins:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QI (SVM)</td>
<td>100 (77–100)</td>
<td>70 (52–84)</td>
<td>56 (35–75)</td>
<td>78 (64–89)</td>
<td>94 (83–98)</td>
</tr>
<tr>
<td>SFU grade 3</td>
<td>100 (77–100)</td>
<td>56 (23–53)</td>
<td>39 (23–56)</td>
<td>58 (41–70)</td>
<td>78 (62–89)</td>
</tr>
<tr>
<td>HI</td>
<td>100 (77–100)</td>
<td>47 (30–64)</td>
<td>42 (25–61)</td>
<td>62 (47–75)</td>
<td>78 (55–88)</td>
</tr>
<tr>
<td><strong>T1/2 greater than 40 mins:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QI (LOG)</td>
<td>100 (63–100)</td>
<td>74 (58–86)</td>
<td>42 (20–66)</td>
<td>78 (64–88)</td>
<td>94 (81–98)</td>
</tr>
<tr>
<td>SFU grade 3</td>
<td>100 (63–100)</td>
<td>33 (20–50)</td>
<td>22 (10–39)</td>
<td>44 (30–59)</td>
<td>88 (56–90)</td>
</tr>
<tr>
<td>HI</td>
<td>100 (63–100)</td>
<td>62 (46–76)</td>
<td>33 (16–55)</td>
<td>68 (53–80)</td>
<td>80 (46–93)</td>
</tr>
</tbody>
</table>

For each case (T1/2 greater than 20, 30 or 40 minutes) SFU grade that resulted in 100% sensitivity was used. Optimized threshold of HI was used for prediction of T1/2 greater than 20, 30 or 40 minutes. Accuracy equals true positive + true negative/total positive + total negative. Due to limited number of cases, exact 95% confidence interval is provided (using binomial distribution).
cannot answer those questions, but it can provide more objective criteria from ultrasonographic imaging that could support prospective studies of congenital obstruction. 

Objective prediction of clinically significant hydronephrosis in children has been an ongoing challenge. Erickson et al. reported that SFU grade 3 has never been associated with reduced differential uptake in pediatric kidneys. However, the dividing line between SFU grades 3 and 4 is not clearly defined or reproducible. A recent report by Dias et al. presented the use of renal pelvic dilatation as highly predictive of the need for surgical intervention, based on decreased function on dimercaptosuccinic acid scanning and not DR. This approach used fetal and postnatal anteroposterior diameter of the renal pelvis, with thresholds of 18 mm for fetal and 16 mm for postnatal measures. This approach was not statistically different from SFU selection using grades 3 and 4, which we have shown performs significantly less well than the QI approach presented in this report. The criterion for the study by Dias et al. was need for pyeloplasty based on nonstandard criteria, rather than defining a threshold for more precise imaging.

Several investigators have measured renal parenchymal area to assess renal status in vesicoureteral reflux and hydronephrosis, and recently in posterior urethral valves. Renal parenchymal volume has been measured using 3-D ultrasound to permit estimation of relative function. These approaches have focused primarily on renal function rather than severity of obstruction, yet indicate the potential usefulness of ultrasound in estimating relative renal function.

To our knowledge, this is the first study to relate renal morphological features from US with clinical parameters of DR. In a previous study proposed by Kim et al. a clinical cohort of 46 patients was evaluated. However, no study of the power of the proposed HI as a meaningful indicator of renal function was presented. We used a data cohort of 50 patients, which provided a retrospective power value of greater than 0.99.

There are several limitations to this pilot study that will be addressed in future research. Despite the satisfactory power value provided by the post hoc power analysis, the number of patients may be relatively small for a condition that is as heterogeneous as hydronephrosis. Even so, the errors noted are small, and if the parameters are conservatively defined, there is less risk of an error of omission (ie missing detection of a patient with an increased $T_{1/2}$). The current methodology requires hand drawn segmentation. While this is not difficult, it is tedious and requires some training, and represents a potential source of interobserver variability. Ultimately the goal is to develop an automatic segmentation algorithm, which would permit rapid evaluation of an image to yield a “hydronephrosis score” that would be available to the clinician for further decision making.

This study does not assess the potential of using this methodology for clinical followup. It is unknown if the US parameters remain consistent or reflect alterations in either direction as hydronephrosis evolves. Hafez and Ross et al. observed that increasing hydronephrosis is indicative of obstruction. However, not all obstructed kidneys exhibit increasing dilatation on followup, and it is unknown if persisting hydronephrosis is safe. It remains to be seen if QI may be able to more readily identify these patients. That area is the subject of an ongoing study.

Table 3. Decrease in MAG3 scanning possible using QI vs SFU grade

<table>
<thead>
<tr>
<th>Threshold</th>
<th>QI</th>
<th>SFU Grade 3</th>
<th>SFU Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{1/2}$ greater than 20 mins:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. MAG3 scans</td>
<td>21</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>% Reduction from cohort</td>
<td>58</td>
<td>29</td>
<td>64</td>
</tr>
<tr>
<td>% Missed critical cases</td>
<td>0</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>$T_{1/2}$ greater than 30 mins:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. MAG3 scans</td>
<td>25</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>% Reduction from cohort</td>
<td>50</td>
<td>28</td>
<td>64</td>
</tr>
<tr>
<td>% Missed critical cases</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>$T_{1/2}$ greater than 40 mins:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. MAG3 scans</td>
<td>19</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>% Reduction from cohort</td>
<td>62</td>
<td>28</td>
<td>64</td>
</tr>
<tr>
<td>% Missed critical cases</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

Figure 3. ROC curves for quantitative imaging (SVM and LOG methods), SFU grading and hydronephrosis index methods for selecting patients with $T_{1/2}$ greater than 30 minutes. Sens., sensitivity. Spec., specificity. Acc., accuracy.
CONCLUSIONS
We have demonstrated the potential clinical usefulness of a semiautomated means of quantitating ultrasonographic images of hydronephrotic kidneys. This approach can decrease the number of diuretic renograms performed by up to 62%, depending on selection criteria used. While diuretic renography remains the gold standard, the described technique has the potential to simplify and enhance our care of children with hydronephrosis.

REFERENCES

EDITORIAL COMMENT
This report is a promising contribution to the never-ending discussion among pediatric urologists about the proper diagnosis of congenital hydronephrosis with regard to a conservative approach, followup and indications for surgical intervention. The semiquantitative approach, based
on ultrasonographic images, may lead to more objective and, perhaps most importantly, reproducible results of diagnosis and outcome measures. So far, semiquantitative US has been related to $t_{1/2}$ on MAG3 renography, a quite controversial parameter, especially in younger children and infants. To some extent, this method is observer dependent but the same is the case in assessment of renography. Future studies will hopefully give us the potential to assess differential renal function using 3-D quantitative imaging as a noninvasive tool without the exposure to radiation.

L. Henning Olsen
Department of Urology, Section of Pediatric Urology
Aarhus University Hospital
Aarhus, Denmark