Interobserver Reliability of Urine Sediment Interpretation

Ron Wald,*† Chaim M. Bell,‡∥§ Rosane Nisenbaum,∥§ Samuel Perrone,∥ Orfeas Liangos,∥ Andreas Laupacis,∥∥§ and Bertrand L. Jaber∥

Divisions of *Nephrology and †General Internal Medicine, St. Michael’s Hospital and University of Toronto, ‡Keenan Research Centre in the Li Ka Shing Knowledge Institute of St. Michael’s Hospital, ∥Department of Health Policy, Management, and Evaluation, University of Toronto, and ∥∥Center for Research on Inner City Health, St. Michael’s Hospital, Toronto, Ontario, Canada; and ∥§Division of Nephrology, Caritas St. Elizabeth’s Medical Center and Tufts University School of Medicine, Boston, Massachusetts

Background and objectives: Urine sediment interpretation is frequently used in the evaluation of patients with kidney disease. There has been no systematic evaluation of the reliability of this diagnostic maneuver.

Design, setting, participants, & measurements: Digital photographs of urine sediment images were acquired from 165 consecutive patients being evaluated by the nephrology consultation service at a tertiary care hospital. Urine sediment images of 100 patients were randomly selected; 86 patients had images that were deemed to be of sufficient quality, and one image per patient was chosen for inclusion in an internet-based questionnaire. For each image, the presence or absence of 14 potential urinary structures was ascertained. Ten nephrologists (senior readers [n = 3]: >10 yr of experience; intermediate readers [n = 3]: 1 to 10 yr of experience; and junior readers [n = 4]: first year of practice) completed the questionnaire. For each urinary structure, we measured the rate of complete agreement among the readers as well as the χ statistic as a marker of agreement beyond chance.

Results: Unanimous agreement was highest (79.1%) regarding the presence of broad and fatty casts and poorest (31.4%) for the identification of dysmorphic red blood cells and white blood cells. Interobserver agreement was best for squamous epithelial cells (χ = 0.54) and hyaline casts (χ = 0.52) and worst for transitional epithelial cells (χ = 0.14) and fatty casts (χ = 0.06). When assessed within strata of physician experience, interobserver agreement was not associated with seniority.

Conclusions: Nephrologists achieved slight to moderate agreement in the identification of structures that are commonly observed in the urine sediment.


Materials and Methods

Generation of a Urine Sediment Image Database

Between June 2005 and January 2006, we constructed a database of images collected from consecutive urine sediments reviewed by members of the Nephrology Consult Service at Tufts Medical Center (Boston, MA). Each patient for whom the service reviewed the urine sediment as part of the clinical workup was eligible for the study. The study was approved by the institutional review board at Tufts Medical Center.

Physicians who evaluated the sediment were encouraged to minimize the time between collection and processing. Fresh urine sediments were spun at 3000 rpm for 5 min. The supernatant was discarded, and the pellet was resuspended. Inspections of the unstained sediment were conducted at low and high power (×400) using an Olympus BX40 microscope (Olympus America, Center Valley, PA). For each urine sediment, a minimum of three high-power photographs were taken using a SPOT Insight camera (Model 3.2.0; Diagnostic Instruments, Sterling Heights, MI). The digital images were then labeled and stored as graphics interchange format (gif) files in a password-protected database.

Development and Administration of a Urine Sediment Questionnaire

Urine sediments from 165 consecutive hospitalized inpatients were collected for the study. A computer-generated program selected 100
study identification numbers that were in turn linked to patients whose urine images were considered for inclusion in the study. Fourteen patients had images that were deemed to be of poor quality by two of the investigators (R.W. and S.P.), and these were excluded from further analysis. For each of the 86 remaining patients, the best quality image was selected for inclusion in the electronic questionnaire. Supplemental Figure 1 displays an image that was included in the questionnaire, whereas Supplemental Figure 2 is an image that was rejected because of poor quality.

The 86 study images were incorporated into an internet-based questionnaire that was distributed to 10 nephrologists working at seven different hospitals in Canada and the United States. The nephrologists reading the images were classified according to years in practice since completion of their nephrology training: Three senior readers had >10 yr of clinical practice, three intermediate readers had between 1 and 10 yr of clinical practice, and four junior readers had completed nephrology training within the previous year. All readers were asked to complete the questionnaire independently and were blinded to the clinical data. For each image, the reader was asked to indicate whether each of 14 possible structures was present or absent by clicking “yes” or “no,” respectively. The structures of interest were red blood cells (RBCs), dysmorphic RBCs, white blood cells (WBCs), renal tubular epithelial cells, transitional epithelial cells, squamous epithelial cells, hyaline casts, fine granular casts, coarse granular/muddy brown casts, RBC casts, WBC casts, renal tubular epithelial cell casts, broad casts, and fatty casts. To ensure complete responses, readers could not proceed to the next image without providing a response to each of the 14 questions for a given image.

Statistical Analysis

Inter-reader reliability was evaluated at the level of the urinary structure, and each structure was assessed independent of the others. Every structure was evaluated on each of the 86 images in the questionnaire.

The degree of inter-reader agreement was calculated within each level of reader seniority and for the 10 readers combined and estimated by the percentage of complete agreement and the $\kappa$ statistic, respectively. Complete agreement was indicated when all readers agreed on the presence or absence of a urinary structure. The $\kappa$ statistic was used to evaluate the degree of agreement above and beyond chance and was categorized as slight ($\kappa = 0.00$ to 0.20), fair ($\kappa = 0.21$ to 0.40), moderate ($\kappa = 0.41$ to 0.60), substantial ($\kappa = 0.61$ to 0.80), and almost perfect ($\kappa > 0.80$) (6).

Our sample size of 86 was sufficient to provide 95% confidence intervals of 10% for levels of agreement between 40 and 80%. Continuous variables are presented as means ± SD and categorical variables as percentages. All analyses were performed using SAS 9.1.3 (SAS Institute, Cary, NC).

Results

Patient Characteristics

Of the 86 patients whose urine sediment images were included in the questionnaire, 40% were women. Mean (± SD) age was 59 ± 15 yr, and on the day of urine sediment assessment, the mean serum creatinine was 2.6 ± 1.2 mg/dl (234 ± 96 $\mu$mol/L). The clinical diagnoses of the Nephrology Consult Service after review of the urine sediment were as follows: acute tubular necrosis (42%), prerenal azotemia (30%), chronic kidney disease (6%), glomerulonephritis (5%), acute interstitial nephritis (3%), acute renal allograft rejection (1%), and unknown/unlisted conditions (13%).

Inter-reader Reliability

Aggregate Results for All 10 Readers. Among cellular elements, agreement was best for squamous epithelial cells, and all 10 readers agreed on their presence or absence 68.6% of the time with an associated $\kappa$ score of 0.54 (Table 1). The next highest complete agreement was seen for transitional epithelial cells (54.7%), although this was associated with the lowest $\kappa$ among cellular elements (0.14). For the remainder of cellular structures, complete agreement ranged between 30 and 40% with $\kappa$ scores reflecting fair agreement except for WBCs, for which interobserver agreement was moderate ($\kappa = 0.47$).

Complete agreement for most types of casts was generally better than for most cellular elements with the notable exception of fine and coarse granular casts, for which agreement was 37.2 and 32.6%, respectively. Agreement beyond chance was strongest for hyaline casts ($\kappa = 0.52$) and lowest for fatty casts ($\kappa = 0.06$).

Interobserver Agreement by Level of Experience. The mean rate of complete agreement across all 14 structures was highest for intermediate readers (75.3 ± 13.4%) followed by junior readers (73.6 ± 14.2%) and senior readers (71.2 ± 15.1%). The average $\kappa$ score was highest for junior readers (0.33 ± 0.14) followed by intermediate (0.27 ± 0.19) and senior readers (0.27 ± 0.14). There was no consistent relationship between the experience of the readers and rates of complete agreement and agreement beyond chance.

Discussion

Although the urine sediment is regarded as an important tool in the nephrologist’s diagnostic armamentarium, this study demonstrates that nephrologists generally achieve only fair to moderate levels of agreement in the identification of important structures. In addition, nephrologists’ level of experience did not correlate with enhanced agreement.

Although we asked nephrologists to comment on the presence of 14 structures potentially seen in the urine sediment, the diagnostic importance of each of these structures is not the same. For example, RBC and WBC casts are hallmarks of relatively rare conditions (proliferative glomerulonephritis and interstitial nephritis, respectively) for which specific renal-protective treatments may be necessary. In view of this, missing these structures under light microscopy may lead the clinician down the wrong diagnostic and therapeutic pathways with important implications for patient care. It is reassuring that complete agreement was relatively high for these structures (approximately 80%) in each stratum of experience. Conversely, $\kappa$ scores were only in the fair range, likely as a result of the low prevalence of these structures and the conditions underlying them (7).

Although the reliability of several common diagnostic maneuvers has been rigorously evaluated (8–10), this is the first study to evaluate systematically the reliability of urine sediment interpretation by nephrologists. Two studies of non-nephrologists have highlighted the challenges associated with urine sediment interpretation. Hillborne et al. (11) documented the limited accuracy of urine sediment interpretation by general medical trainees, and another study demonstrated that knowl-
edge of the patient’s clinical details had the potential to bias interpretation of the urine sediment (12). Conversely, one study indicated that urine sediment interpretation by a nephrologist who was blinded to the clinical data yielded a diagnosis that coincided with that of the treating physicians on 90% of occasions (13). It was also suggested that urine sediment reports emanating from the clinical laboratory may be inaccurate and lead to incorrect diagnoses; however, there was no inter-nephrologist comparison with respect to the identification of urine sediment structures.

Strengths of this study were the use of a standardized electronic questionnaire and blinding of readers to the clinical data. Our archive of urine sediment photographs from a series of consecutive patients ensured that the images reflected actual clinical practice.

The limitations of our study require discussion. We studied the responses of only 10 nephrologists. Although the readers all were recognized as skilled clinicians, we cannot be certain that their responses were representative of the wider nephrology community, and our findings should be confirmed among a larger cohort of readers. Although we attempted to standardize the processing and interpretation of the urine sediment, the time elapsed between bedside urine collection and microscopic interpretation was likely variable. Nonuse of phase contrast microscopy and polarized light limited the readers’ ability to comment optimally on the presence of dysmorphic RBCs (14) and fatty casts (15). Readers were also presented with fixed images that were not amenable to fine focusing or navigation through multiple image fields, as is feasible when using a microscope. Our questionnaire was relatively long and repetitive (86 images with 14 questions per image totaling 1204 responses per reader), which may have compromised a reader’s ability to maintain consistent attention; however, readers had the option of stopping the questionnaire at anytime and resuming at their leisure. Because of the lack of a clear reference standard, it was difficult to ascertain the true prevalence for each of the structures that were studied. Given the relative rarity of conditions such as acute interstitial nephritis and glomerulonephritis in everyday practice, it is likely that WBC and RBC casts were rare in our image set. Because it is widely

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prevalence (n = 86)</th>
<th>Senior Readers (n = 3)</th>
<th>Intermediate Readers (n = 3)</th>
<th>Junior Readers (n = 4)</th>
<th>All Readers (n = 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Complete Agreement (%)</td>
<td>Complete Agreement (%)</td>
<td>Complete Agreement (%)</td>
<td>Complete Agreement (%)</td>
</tr>
<tr>
<td>RBC</td>
<td>56(65.1%)</td>
<td>50.0 0.24</td>
<td>68.6 0.41</td>
<td>75.6 0.30</td>
<td>39.5 0.29</td>
</tr>
<tr>
<td>dysmorphic RBC</td>
<td>24(27.9%)</td>
<td>44.2 0.14</td>
<td>54.7 0.13</td>
<td>75.6 0.33</td>
<td>31.4 0.24</td>
</tr>
<tr>
<td>WBC</td>
<td>25(29.1%)</td>
<td>67.4 0.49</td>
<td>65.1 0.50</td>
<td>48.8 0.42</td>
<td>31.4 0.47</td>
</tr>
<tr>
<td>renal tubular epithelial cell transitional epithelial cell squamous epithelial cell</td>
<td>18(20.1%)</td>
<td>57.0 0.28</td>
<td>66.3 0.28</td>
<td>55.8 0.21</td>
<td>32.6 0.20</td>
</tr>
<tr>
<td></td>
<td>14(16.3%)</td>
<td>62.8 0.21</td>
<td>84.9 0.02</td>
<td>75.6 0.12</td>
<td>54.7 0.14</td>
</tr>
<tr>
<td></td>
<td>9(10.5%)</td>
<td>83.7 0.49</td>
<td>84.9 0.57</td>
<td>83.7 0.52</td>
<td>68.6 0.54</td>
</tr>
<tr>
<td>Hyaline cast</td>
<td>13(15.1%)</td>
<td>81.4 0.50</td>
<td>82.6 0.61</td>
<td>68.6 0.46</td>
<td>59.3 0.52</td>
</tr>
<tr>
<td>Fine granular cast</td>
<td>9(10.5%)</td>
<td>65.1 0.19</td>
<td>64.0 0.28</td>
<td>65.1 0.35</td>
<td>37.2 0.27</td>
</tr>
<tr>
<td>Coarse granular/ muddy brown cast</td>
<td>26(30.2%)</td>
<td>59.3 0.31</td>
<td>52.3 0.20</td>
<td>48.8 0.15</td>
<td>32.6 0.22</td>
</tr>
<tr>
<td>RBC cast</td>
<td>2(2.3%)</td>
<td>82.6 0.22</td>
<td>79.1 0.25</td>
<td>83.7 0.39</td>
<td>62.8 0.31</td>
</tr>
<tr>
<td>WBC cast</td>
<td>2(2.3%)</td>
<td>86.0 0.09</td>
<td>83.7 0.16</td>
<td>84.9 0.39</td>
<td>73.3 0.27</td>
</tr>
<tr>
<td>Renal tubular epithelial cell cast</td>
<td>6(6.7%)</td>
<td>77.9 0.21</td>
<td>82.6 0.15</td>
<td>90.7 0.57</td>
<td>65.1 0.26</td>
</tr>
<tr>
<td>Broad cast</td>
<td>3(3.5%)</td>
<td>87.2 0.27</td>
<td>90.7 0.17</td>
<td>86.0 0.35</td>
<td>79.1 0.32</td>
</tr>
<tr>
<td>Fatty cast</td>
<td>1(1.7%)</td>
<td>91.9 0.10</td>
<td>95.3 −0.02</td>
<td>87.2 0.11</td>
<td>79.1 0.06</td>
</tr>
</tbody>
</table>

*aPrevalence was defined as the proportion of images on which the structure was observed by at least 2 of the senior readers. Complete agreement was defined as the percentage of images where all readers in that stratum of experience agreed that the structure was present or absent. RBC, red blood cell; WBC, white blood cell.
known that the $\kappa$ value is often low when a finding is rare (7), some of the low $\kappa$ statistics must be interpreted with caution. Given the presumed variability in the prevalence of the different structures in the various images, comparison of $\kappa$ scores across urinary structures should be performed cautiously.

Difficulties in the establishment of a reference standard for urine sediment interpretation limited our ability to assess accuracy (16). Unlike other physical examination maneuvers whereby readily obtainable tests exist to objectively validate a clinical finding (e.g., echocardiography to establish valvular pathology), no parallel exists for urine sediment interpretation. A more useful reference standard would be a kidney biopsy; however, this is not possible for the purposes of a study because of the risks of the procedure. Moreover, findings from a kidney biopsy would not account for structures emanating from the lower urinary tract.

The limited reliability of urine sediment interpretation by nephrologists raises important questions about the role of this test in the evaluation of patients with kidney disease. Our results can be used to support nephrologists who have abandoned urinalysis or perhaps justify further efforts to automate the interpretation of the urine sediment (17–21); however, we believe that the ready availability and low cost of urine examination and the potential to derive important diagnostic data justify efforts to optimize the quality of urine sediment interpretation by clinicians.

If nephrologists are to maintain urine sediment analysis as an integral part of their practice, then we propose several solutions. Urine sediment interpretation should become more standardized, with clear criteria to define the characteristics of each structure, a practice that has been adopted in other settings (22). Such initiatives should lead to the unification of the terminology used by those interpreting urine sediments. This should be complemented by education initiatives that are spearheaded by bodies that have called for proficiency in urinalysis and the potential to derive important diagnostic data justify efforts to optimize the quality of urine sediment interpretation by clinicians.

Conclusions

The inter-reader agreement of urine sediment interpretation by nephrologists is disappointing. To justify the continued use of the urine sediment in the diagnosis of kidney disease, measures to standardize its interpretation and enhance training for readers are needed.

Acknowledgments

R.W. is supported by the Canadian Institutes of Health Research and has received an unrestricted fellowship award from Amgen. C.M.B. is the recipient of a Canadian Institutes of Health Research, Institute of Aging New Investigator Award. R.N. gratefully acknowledges the support of the Ontario Ministry of Health and Long-Term Care. B.L.J. is supported by grants from the National Institutes of Health (DK065102 and DK077751).

We are grateful for the time and efforts of the nephrologists who interpreted the urine sediment images. Other individuals who made this study possible include the nephrology fellows and staff at Tufts Medical Center, who assembled the database of sediment images: Dr. Ronald Perrone for assisting with the urine photography and troubleshooting of the equipment; Dr. Xinogang Gao for creating and managing the web-based questionnaire; and Drs. Sharon Strauss, Joel Ray, and Ahmed Bayoumi for insightful reviews of the manuscript.

Disclosures

None.

References

16. Jaeschke R, Guyatt G, Sackett DL: Users’ guides to the medical literature: III. How to use an article about a diag-
24. Objectives of Training and Specialty Training Requirements in Nephrology, Ottawa, Royal College of Physicians and Surgeons of Canada, 2005

Access to UpToDate on-line is available for additional clinical information at http://www.cjasn.org/

Supplemental information for this article is available online at http://www.cjasn.org.