Research Article

Interobserver reproducibility of the Paris System for Reporting Urinary Cytology

Theresa Long, BS, Lester J. Layfield, MD*, Magda Esebua, MD, Shellaïne R. Frazier, DO, D. Tamar Giorgadze, MD, PhD¹, Robert L. Schmidt, MD, PhD, MBA²

Address: Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, Missouri, ¹Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, ²Department of Pathology and Laboratory Medicine and ARUP Laboratories, University of Utah, Salt Lake City, Utah, USA

E-mail: Theresa Long - tmlgx2@health.missouri.edu; Lester J. Layfield* - layfieldl@health.missouri.edu; Magda Esebua - esebuam@health.missouri.edu; Shellaïne R. Frazier - fraziersr@health.missouri.edu; D. Tamar Giorgadze - tgiorgadze@mcw.edu; Robert L. Schmidt - robert.schmidt@hsc.utah.edu

*Corresponding author

INTRODUCTION

The diagnosis of urothelial carcinoma and posttherapy follow-up requires examination of either histologic biopsy specimens or cytologic examination of voided urine or instrumented urinary tract specimens. The diagnosis of urothelial carcinoma by cytologic methods
can be challenging, especially the recognition of low-grade urothelial carcinomas. A number of diagnostic approaches and classification schemes have been proposed over the last seven decades.[1-9] Diagnostic classifications and criteria have been proposed by Papanicolaou,[1] Koss et al.[2] Murphy et al.,[3] Ooms and Veldhuizer,[4] and The Papanicolaou Society of Cytopathology.[5] All of these systems have been highly successful for the recognition of high-grade urothelial carcinomas (HGUCs) but have had low sensitivity for the recognition of low-grade urothelial carcinomas.[1-3] A number of technologies have been brought to bear to improve diagnostic sensitivity of urinary cytology,[6,7] but the success of these methodologies has often been insufficient to justify their cost. Fluorescence in situ hybridization analysis has been one of the more popular, but some studies have shown that its sensitivity for low-grade papillary urothelial carcinoma is not significantly superior to that of cytology alone.[5]

The recently proposed Paris System for Reporting Urinary Cytology advocates a classification to improve the sensitivity and specificity for the diagnosis of HGUC.[10] This system utilizes seven categories designated: (1) inadequate/less than optimal adequacy; (2) negative for high-grade urothelial carcinoma (NHGUC); (3) atypical urothelial cells (AUC); (4) low-grade urothelial neoplasm (LGUN); (5) suspicious for high-grade urothelial carcinoma (SHGUC); (6) HGUC; and (7) other malignancies. Each of these categories is well defined and classified as unsatisfactory or other). Statistical calculations were performed using Stata 14 (StataCorp., College Station, Texas, USA). Weighted averages were calculated using the MetaProp command. The average kappa statistics were calculated using the Kapci command.

Currently, little published data exist documenting the interobserver reproducibility of these categories. For clinical utility, a categorization scheme must be both accurate and precise. We investigated the interobserver reproducibility of five categories used in The Paris System for Reporting Urinary Cytology. The analysis was performed by four cytopathologists who had not participated in the development of The Paris System for Reporting Urinary Cytology. Herein, we report the results of our reproducibility study for evaluation of precision of the Paris System.

**METHODS**

The study design was reviewed by the Institutional Review Board at the University of Missouri for compliance with university, national, and international standards. The Institutional Review Board designated the study as "exempt." Three hundred and fifty-seven urinary cytology specimens (328 voided urines, 13 catheterized urines, 10 ureteral brushings and washings, 4 obtained during cystoscopy, and 2 obtained from the kidney) obtained over a 10-year period were selected for the study. Only cases with well-fixed, well-prepared, liquid-based preparations were chosen for inclusion in the study. The majority of cases were voided urines. All specimens were Papanicolaou-stained ThinPrep preparations. Each case was reviewed independently by four cytopathologists unaware of the previous diagnoses and unaware of the diagnoses given by other cytopathologists participating in the study. The cytopathologists had between 6 and 25 years’ experience with interpreting urinary cytology. The categories used for assignment were those of the Paris System and included: unsatisfactory, NHGUC, AUC, LGUN, SHGUC, HGUC, and other malignancies.[10] The four review cytopathologists had not been involved in the development of The Paris System for Reporting Urinary Cytology, but each cytopathologist read the monograph entitled, The Paris System for Reporting Urinary Cytology and applied the criteria as outlined in the relevant chapters (3, 4, 5, 6, and 7).[6] The nuclear-cytoplasmic (N/C) ratio was estimated visually using the definitions proffered in the Paris System monograph [Table 1 for criteria used]. One of the cytopathologists had attended lectures at national meetings outlining the Paris System. The category LGUN was recognized as a subcategory of NHGUC and was only used where papillary groups of urothelial cells were present with well-defined fibrovascular cores.

Overall agreement was assessed using absolute agreement and weighted chance-corrected agreement (kappa). For weighted kappa, concordant results were given full credit (1) and discordant results were given half credit if the discordance was off by a single category [Table 2]. For kappa calculations, the categories were ordered as follows: NHGUC greater than AUC greater than LGUN greater than SHGUC greater than HGUC. The categories of "unsatisfactory" and "other malignancy" were excluded from the agreement analysis. The overall average absolute agreement and expected agreement were calculated by determining the agreement between each pair of observers and calculating the weighted average (each pair had a different number of cases due to exclusion of cases classified as unsatisfactory or other). Statistical calculations were performed using Stata 14 (StataCorp., College Station, Texas, USA). Weighted averages were calculated using the MetaProp command. The average kappa statistics were calculated using the Kapci command.

To further characterize agreement, we investigated the reliability of specific rating categories. To that end, we tabulated the results of all the rater comparisons and determined the frequency of each possible combination of ratings (NHGUC vs. AUC, NHGUC vs. SHGUC, etc.). We categorized each combination as high clinical impact, low clinical impact, or no impact according to the potential impact of a misclassification on clinical management. We determined the reliability of individual categories by
RESULTS AND DISCUSSION

The diagnoses were unevenly distributed across categories [Table 3]. NHGUC (44.5%) and AUC (22%) were the most common diagnoses. Of 1902 comparisons between rater pairs, 999 (52.5%) showed complete agreement, 451 (23.7%) differed by a single category, 228 (12%) differed by two categories, 199 (10.5%) differed by three categories, and 25 (1.3%) differed by four categories. Using the impact categories in Table 4, 14.7% of the disagreements were considered “high clinical impact.” The average absolute agreement was 65% (95% confidence interval: 63–67). The average expected agreement (due to chance alone) was 44% (95% confidence interval: 42–46). The average chance-corrected agreement (kappa) was 0.32 (95% confidence interval: 0.28–0.32).

The agreement by category is shown in Table 5. Raters had the highest agreement (63%) on NHGUC. Raters showed poorer agreement on the other categories. For example, given a diagnosis of SHGUC by one rater, the chance of agreement was only 13% and the second diagnoses were evenly distributed over the other categories. Figures 1-4 show examples of specimens with poor interobserver agreement.

Cytologic analysis of specimens obtained from the urinary tract remains a cornerstone for the diagnosis and surveillance of patients with urothelial carcinoma. The sensitivity and specificity of cytologic analysis are excellent for HGUCs. Cytologic analysis is accurate for both superficial high-grade lesions as well as those extending deeper into the bladder wall. Unfortunately, the sensitivity of cytologic study for the diagnosis of LGUCs is significantly poorer. In addition, noncarcinomatous papillary lesions exist (papilloma and papillary urothelial lesion of low malignant potential) complicating the cytologic diagnosis of low-grade urothelial carcinomas. The Papanicolaou Society of Cytopathology Task Force recommendations acknowledged this issue and included the category “atypical urothelial cells.” More recently, a number of investigators have studied the clinical significance of the atypical category and attempted to define useful morphologic criteria to further stratify malignancy risk. Published data are inconclusive as to whether specimens designated “atypical” have increased risk over those designated “negative.”[9,10] Cytologic features have been identified which correlate with increased risk for low-grade urothelial carcinoma.[11,12] Previously, many cytopathologists subdivide the atypical category into “atypical, suspicious for malignancy” and “atypical, favor reactive.”[9,10] Such a subdivision may increase diagnostic accuracy, but the malignancy risk of these two categories and their reproducibility are incompletely understood. The relationship between the

Table 1: Criteria for assignment to the diagnostic categories of the Paris System for Reporting Urinary Cytology

| NHGUC (negative) |
|------------------|
| A sample of urine, either voided or instrumented, may be considered negative (NHGUC) if any of the following components is present in the specimen |
| Benign urothelial, glandular, and squamous cells |
| Benign urothelial tissue fragments and urothelial cell clusters |
| Changes associated with lithiasis |
| Viral cytopathic effect |
| Posttherapy effect, including epithelial cells from urinary diversions |
| AUC: Atypical urothelial cells, N/C: Nuclear to cytoplasmic, SHGUC: Suspicious for high-grade urothelial carcinoma, NHGUC: Negative for high-grade urothelial carcinoma |
| HGUC: High-grade urothelial carcinoma, LGUN: Low-grade urothelial neoplasm, AUC: Atypical urothelial cells, N/C: Nuclear to cytoplasmic, SHGUC: Suspicious for high-grade urothelial carcinoma, NHGUC: Negative for high-grade urothelial carcinoma |

Table 2: Weighting scheme for agreement analysis

| NHGUC | AUC | LGUN | SUS | HGUC |
|-------|-----|------|-----|------|
| 1     |     |      |     |      |
| 0.5   | 1   |      |     |      |
| 0     | 0.5 | 1    |     |      |
| 0     | 0   | 0.5  | 1   |      |
| 0     | 0   | 0    | 0.5 | 1    |

Concordant results were given full credit. Discordant results were given half credit if the discordance was off by one category. LGUN: Low-grade urothelial neoplasm, AUC: Atypical urothelial cells, NHGUC: Negative for high-grade urothelial carcinoma, SUS: Suspicious, HGUC: High-grade urothelial carcinoma.

determining the conditional probability of a disagreement by a second rater, given a particular rating by the first rater.
category “atypical suspicious for malignancy” and the presence of low-grade or HGUC is also poorly defined.

Recently, The Paris System for Reporting Urinary Cytology was developed to improve diagnostic accuracy and clarify the usage and implications of the atypical category or categories. This system effectively has two atypical categories named as: (1) AUCs and (2) SHGUC. The categories AUCs and SHGUC are defined by cytologic features including the N/C ratio. Recently, Layfield et al. demonstrated imperfect interobserver reproducibility of estimates of N/C ratio, and Zhang et al. demonstrated that morphologists overestimated the N/C ratio. The category LGUN is well defined by both cellular and architectural features. The category LGUN should have limited usage and is considered a subcategory of NHGUC. While the categories of The Paris System have good estimates of malignancy risk, we are unaware of any study documenting the interobserver reproducibility of these categories. Such data are needed to estimate the precision of The Paris System for Reporting Urinary Cytology.

Our study was designed to evaluate diagnostic precision not diagnostic accuracy of The Paris System for Reporting Urinary Cytology, and thus, follow-up histologic diagnoses were not necessary. We investigated the reproducibility of category assignment by four cytopathologists reviewing 357 specimens. The clinical utility of any classification system depends on both its precision (reproducibility) and its diagnostic accuracy. Prior classifications of urinary tract specimens were often imprecise in relation to the term “atypical” and for the diagnosis of low-grade urothelial carcinomas. The frequency of use of the term “atypical” has ranged from 2% to 31%, and the risk of malignancy associated with the term has varied from 8.3% to 37.5%. Hence, use of the term varies widely as does its relationship to the presence of malignancy. Recently, The Paris System for Reporting Urinary Cytology was developed to improve both precision and accuracy for the diagnosis

Table 3: Tabulated scores for each observer

| Score   | Observer | Average (%) |
|---------|----------|-------------|
| USAT    | 9        | 6 (1.7)     |
| NHGUC   | 121      | 158 (44.5)  |
| AUC     | 85       | 78 (22.0)   |
| LGUN    | 3        | 15 (4.2)    |
| SUS     | 70       | 40 (11.3)   |
| HGUC    | 31       | 43 (12.1)   |
| Other   | 38       | 15 (4.2)    |
| Total   | 357      |             |

NHGUC: Negative for high-grade urothelial carcinoma, AUCs: Atypical urothelial cells, LGUN: Low-grade urothelial neoplasm, HGUC: High-grade urothelial carcinoma, SUS: Suspicious, USAT: Unsatisfactory

Table 4: Overall agreement between all pairs of observers

| Disagreement | Cases | Percentage | Size of disagreement | Potential clinical impact |
|--------------|-------|------------|----------------------|--------------------------|
| None         | 999   | 52.5       | 0                    | None                     |
| NHGUC versus AUC | 270 | 14.2       | 1                    | Low                      |
| AUC versus LGUN | 51   | 2.7        | 1                    | Low                      |
| SUS versus LGUN | 22   | 1.2        | 1                    | Low                      |
| SUS versus HGUC | 108  | 5.7        | 1                    | Low                      |
| NHGUC versus LGUN | 15  | 0.8        | 2                    | Low                      |
| AUC versus SUS | 155  | 8.1        | 2                    | Low                      |
| LGUN versus HGUC | 58   | 3.0        | 2                    | High                     |
| NGUN versus SUS | 90   | 4.7        | 3                    | High                     |
| AUC versus HGUC | 109  | 5.7        | 3                    | High                     |
| NHGUC versus HGUC | 25   | 1.3        | 4                    | High                     |
| Total        | 1902  | 100.0      |                      |                          |

For example, out of 1902 comparisons, 999 showed complete agreement. The size of disagreement indicates the number of categories between two diagnoses. NHGUC: Negative for high-grade urothelial carcinoma, AUCs: Atypical urothelial cells, LGUN: Low-grade urothelial neoplasm, HGUC: High-grade urothelial carcinoma, SUS: Suspicious, USAT: Unsatisfactory.
of urothelial carcinoma. Definitive diagnostic criteria were proposed. While malignancy risk was associated with each category, little information is available on the precision of assignment of specimens to these diagnostic categories. Brimo and Auger [21] pointed out that even when the “atypical” category is precisely defined and accurately applied, variability in the incidence of the “atypical” diagnosis and its association with carcinoma will vary between laboratories and patient populations studied. Our study reported an AUC rate of 22%. While this is high in comparison to some reports,[17,19,22] it is less than the AUC rates reported by Rosenthal et al.[23] (31%) and Brimo and Auger[21] (26%). Our finding of an AUC rate of 22% continues to demonstrate the difficulty associated with classification of some urine specimens despite the more precise definitions of The Paris System. The atypical cells of uncertain significance (ACUS) category also presents challenges for the Bethesda system for cervical cytology. This was addressed by adding categories ACUS high and ACUS low, which do not necessarily improve precision but appear to have increased the clinical utility of the Bethesda system for cervical cytology. Despite the poor precision of the category, AUCs, this category appears to have clinical value because it stratifies malignancy risk. Because of the clinical value of the atypical category, more reproducible criteria are desirable. While the N/C ratio would appear to be measurable and reproducible, it appears imperfect in clinical practice and in a number of morphometric studies.[14] A potential problem with the Paris System for urinary cytology is not the individual criteria including N/C ratio but the propensity for cytopathologists to use atypical categories as a “wastebasket” for cases where a cytopathologist is not certain a specimen is benign but does not see sufficient changes to be comfortable with a suspicious for malignancy diagnosis. In some of these cases just calling a specimen atypical may seem like a safe diagnostic option for some cytopathologists despite the published criteria. Large linear regression studies may be helpful in developing more reproducible criteria for the atypical category.

We assessed clinical impact of disagreement for the diagnostic pairs as shown in Table 4. In the majority of diagnostic pairings, the diagnoses varied by only one category and the clinical impact of such a disagreement between observers would be low with little effect on management. However, disagreement by two or more categories could have a significant impact on clinical management. Hence, classification discrepancy pairs or “NHGUC vs. SHGUC,” “AUC vs. HGUC,” “LGUN vs. HGUC,” and “NHGUC vs. HGUC” could have a significant impact on how a patient is clinically managed. Such “high impact” disagreement occurred in approximately 15% of cases [Table 4]. The probability that a second rater would agree with the initial rater varied considerably [Table 5].
Agreement for the category NHGUC was good at 63%. Agreement for the other categories was poor, especially for the categories LGUN (6%) and SHGUC (13%). Even the categories AUC and HGUC had relatively poor agreement between raters (21%). The Paris System for Reporting Urinary Cytology demonstrates good precision for the diagnosis of NHGUC but had poor interobserver reproducibility (precision) for the other categories. This finding is similar to issues identified with prior categorization proposals. Kassouf observed that the category “atypical urothelial cells” is frequently reported raising concerns regarding the possible presence of cancer which may lead to over investigation and excessively prolonged surveillance. The new Paris System may reduce the percentage of urinary tract samples designated as “atypical.” Our series reported an AUS rate of 22%, which is lower than that found by some authors using other classification systems but is still higher than an optimal AUS rate of below 10% as reported by some authorities.

CONCLUSIONS

Clinical outcomes analysis will be necessary to determine the implications of our findings of poor interobserver agreement for the majority of the categories used in The Paris System for Reporting Urinary Cytology. Our findings suggest that the Paris System suffers from an unacceptably low precision in some of its diagnostic categories. Further, large studies will be necessary to demonstrate the precision and accuracy of The Paris System for Reporting Urinary Cytology.

COMPETING INTERESTS STATEMENT BY ALL AUTHORS

The author(s) declare that they have no competing interests.

AUTHORSHIP STATEMENT BY ALL AUTHORS

All authors of this article declare that we qualify for authorship as defined by ICMJE.

Each author has participated sufficiently in the work and take public responsibility for appropriate portions of the content of this article.

TL carried out data collection and organized. LJL developed the concept and wrote the article. ME, SRF, and TG reviewed slides. RLS completed the statistical analysis. All authors read and approved the final manuscript.

ETHICS STATEMENT BY ALL AUTHORS

This study was conducted with approval from Institutional Review Board (IRB) of all the institutions associated with this study as applicable.

Authors take responsibility to maintain relevant documentation in this respect.

LIST OF ABBREVIATIONS (In alphabetic order)

ACUS - Atypical cells of uncertain significance
AUC - Atypical urothelial cells
HGUC - High grade urothelial carcinoma
LGUN - Low grade urothelial neoplasm
N/C - Nuclear/cytoplasmic
NHGUC - Negative for high grade urothelial carcinoma
SHGUC - Suspicious for high grade urothelial carcinoma
USA - United States of America.

REFERENCES

1. Papanicolaou GN. Cytology of the urine sediment in neoplasms of the urinary tract. J Urol 1947;57:375-9.
2. Koss LG, Bartels PH, Sychra JJ, Wied GL. Diagnostic cytologic sample profiles in patients with bladder cancer using TICAS system. Acta Cytol 1978;22:392-7.
3. Murphy WM, Soloway MS, Jukkola AF, Crabtree WN, Ford KS. Urinary cytology and bladder cancer. The cellular features of transitional cell neoplasms. Cancer 1984;53:1555-65.
4. Ooms EC, Veldhuizen RW. Cytological criteria and diagnostic terminology in urinary cytology. Cytopathology 1993;4:51-4.
5. Layfield LJ, Elsheikh TM, Fili A, Nayar R, Shidham V; Papanicolaou Society of Cytopathology. Review of the state of the art and recommendations of the Papanicolaou Society of Cytopathology for urinary cytology procedures and reporting: The Papanicolaou Society of Cytopathology Practice Guidelines Task Force. Diagn Cytopathol 2004;30:24-30.
6. Toma MI, Friedrich MG, Hautmann SH, Jäkel KT, Erbersdobler A, Hellstern A, et al. Comparison of the ImmunoCyt test and urinary cytology with other urine tests in the detection and surveillance of bladder cancer. World J Urol 2004;22:145-9.
7. Moonen PM, Merkx GF, Peelen P, Karthaus HF, Smeets DF, Witjes JA. UroVysion compared with cytology and quantitative cytology in the surveillance of non-muscle-invasive bladder cancer. Eur Urol 2007;51:1275-80.
8. Rosenthal DL, Wojcik EM, Kursycz DF. In: The Paris System for Reporting...
9. Hughes JH, Raab SS, Cohen MB. The cytologic diagnosis of low-grade transitional cell carcinoma. Am J Clin Pathol 2000;114(Suppl 1):S59-67.

10. Renshaw AA, Nappi D, Weinberg DS. Cytology of grade I papillary transitional cell carcinoma. A comparison of cytologic, architectural and morphometric criteria in cystoscopically obtained urine. Acta Cytol 1996;40:676-82.

11. Deshpande V, McKee GT. Analysis of atypical urine cytology in a tertiary care center. Cancer 2005;105:468-75.

12. Brimo F, Vollmer RT, Case B, Aprikian A, Kassouf W, Auger M. Accuracy of urine cytology and the significance of an atypical category. Am J Clin Pathol 2009;132:785-93.

13. Renshaw AA. Subclassifying atypical urinary cytology specimens. Cancer 2000;90:222-9.

14. Layfield LJ, Schmidt RL, Esebua M, Frazier SR, Hammer RD, Bivin WW, et al. Accuracy and reproducibility of nuclear/cytoplasmic ratio assessments: An intra-and interobserver study. Diagn Cytopathol 2017;45:107-12.

15. Zhang ML, Guo AX, VandenBussche CJ. Morphologists overestimate the nuclear-to-cytoplasmic ratio. Cancer Cytopathol 2016;124:669-77.

16. Piaton E, Decaussin-Petrucci M, Mege-Lechevallier F, Advenier AS, Devonec M, Ruffion A. Diagnostic terminology for urinary cytology reports including the new subcategories 'atypical urothelial cells of undetermined significance' (AUC-US) and 'cannot exclude high grade' (AUC-H). Cytopathology 2014;25:27-38.

17. Muus Ubago J, Mehta V, Wojciek EM, Barkan GA. Evaluation of atypical urinary cytology progression to malignancy. Cancer Cytopathol 2013;121:387-91.

18. Bhatia A, Dey P, Kakkar N, Srinivasan R, Nijhawan R. Malignant atypical cell in urine cytology: A diagnostic dilemma. Cytojournal 2006;3:28.

19. Mokhtar GA, Al-Dousari M, Al-Ghamdi D. Diagnostic significance of atypical category in the voided urine samples: A retrospective study in a tertiary care center. Urol Ann 2010;2:100-6.

20. VandenBussche CJ, Sathiyamoorthy S, Owens CL, Burroughs FH, Rosenthal DL, Guan H. The Johns Hopkins Hospital template for urologic cytology samples: Parts II and III. Improving the predictability of indeterminate results in urinary cytologic samples: An outcomes and cytomorphic study. Cancer Cytopathol 2013;121:21-8.

21. Brimo F, Auger M. The atypical urothelial cell category in the Paris system: Strengthening the Achilles’ heel. Cancer Cytopathol 2016;124:305-6.

22. Barasch S, Choi M, Stewart J 3rd, Das K. Significance of atypical category in voided urine specimens prepared by liquid-based technology: Experience of a single institution. J Am Soc Cytopathol 2014;3:118-25.

23. Rosenthal DL, Vandenbussche CJ, Burroughs FH, Sathiyamoorthy S, Guan H, Owens C. The Johns Hopkins Hospital template for urologic cytology samples: Part I-creating the template. Cancer Cytopathol 2013;121:15-20.

24. Kassouf W. The value of urine cytology in the workup of hematuria. Cancer Cytopathol 2016;124:303-4.

EDITORIAL/PEER REVIEW STATEMENT

To ensure the integrity and highest quality of CytoJournal publications, the review process of this manuscript was conducted under a double-blind model (authors are blinded for reviewers and vice versa) through automatic online system.