Grading variability of urothelial carcinoma: experience from a single academic medical center

Eugene W. Lee, MD,[1] Fang-Ming Deng, MD,[2] Jonathan Melamed, MD,[2] Savvas Mendrinos, MD,[2] Kasturi Das, MD,[2] Tsivia Hochman, PhD,[3] Samir S. Taneja, MD,[1] William C. Huang, MD[1]

[1]Department of Urology, New York University Langone Medical Center, New York, New York, USA
[2]Department of Pathology, New York University Langone Medical Center, New York, New York, USA
[3]Department of Biostatistics, New York University Langone Medical Center, New York, New York, USA

LEE EW, DENG F-M, MELAMED J, MENDRINOS S, DAS K, HOCHMAN T, TANEJA SS, HUANG WC. Grading variability of urothelial carcinoma: experience from a single academic medical center. *Can J Urol* 2014;21(4):7374-7378.

Introduction: *Tumor grade plays a critical role in the management of papillary non-invasive urothelial carcinoma (UC). Since grading of UC relies on morphologic criteria, variability in interpretation exists among pathologists. The objective of this study was to examine inter-observer variability in the grading of papillary non-invasive UC at a single academic medical center.*

Materials and methods: *One general pathologist and two genitourinary pathologists were blinded to patient identity and graded 98 consecutive UC specimens using the 1973 and 2004 classification systems. Kappa statistics (κ) were used to measure inter-observer reproducibility to account for agreement expected purely by chance. By convention, κ values from 0.21-0.4 represent "fair", from 0.41-0.6 represent "moderate", and > 0.6 represent "substantial" agreement.*

Results: *Raw percentage agreement among all three pathologists was only 26% using the 1973 system and 47% using the 2004 system. When measured by kappa, overall agreement was only "fair" for both systems and while higher for the 2004 system than the 1973, this was not significant (κ: 0.38 versus 0.26, respectively). There were no significant differences in agreement when comparing the specialists' agreement between themselves with agreement between each specialist and the generalist (κ: 0.31-0.37 versus κ: 0.18-0.46).*

Conclusions: *The current grading system continues to demonstrate challenges in reproducibility among general and specialized pathologists. The degree of variability has significant implications on management decisions for non-invasive UC. Our findings underscore the need to identify molecular markers that can provide a more objective and reliable risk stratification system to guide patient management.*

Key Words: superficial, bladder cancer, urothelial, grade, agreement, reproducibility

Introduction

Urothelial carcinoma (UC) of the bladder is one of the most commonly diagnosed neoplasms, with over 70,000 new patient diagnoses projected in the United States for 2014 contributing to a prevalence of over 500,000 cases.[1] While tumor stage remains the single most important prognostic factor, the majority of patients present with superficial, non-muscle invasive tumors. Within this subset of patients, pathological grade of the tumor is the primary factor used by clinicians to predict progression to advanced disease, and is relied upon heavily when making therapeutic decisions.[2] Pathologists have traditionally used morphological criteria to identify tumor grade. Low grade tumors are prone to recurrence but unlikely to progress, while high grade tumors have a propensity for invasion and metastasis and thus are treated aggressively. There is strong evidence supporting a molecular basis for this divergence in morphology and behavior.[3]

Accepted for publication May 2014

Address correspondence to Dr. William C. Huang, Department of Urology, NYU School of Medicine, 150 East 32nd Street – 2nd Floor, New York, NY 10016 USA

The reproducibility of pathologic grading of bladder tumors has been long called into question. Numerous grading systems exist, but the 3-tier 1973 World Health Organization (WHO) system was the most widely used for decades, in which tumors are graded with increasing anaplasia from 1 to 3. However, due to a lack of distinct criteria defining each grade, it suffered from marked inter-observer variability.[4] The effect is clearly evident from the wide range in reported frequency of grade 2 tumors with incidences ranging from 13% to 69%.[5] This was problematic as there was no generally accepted guideline for management for grade 2 tumors, and therefore clinicians were forced to treat them as either low grade, high grade, or some combination of the two.

In order to create a more universally applicable classification with acceptable reproducibility, a new system was first introduced by members of the International Society of Urological Pathology (ISUP) in 1998, and several years later was adopted as the 2004 WHO/ISUP classification system.[6] Through elimination of grade 2 by designating tumors as either low or high grade, in addition to providing detailed histologic criteria for each grade, the 2004 revision aimed to be a "universally acceptable classification system for bladder neoplasms that could be used effectively by pathologists, urologists, and oncologists". The prognostic utility of the 2004 system has been validated in several retrospective studies, and two recent prospective studies have confirmed this.[7,8]

It is not clear, however, if the revision has had a profound effect on reproducibility of the 1973 system. Several reports have found varying levels of agreement for the 2004 system, and shown only slight or no improvement over the 1973 system.[9-12] However, almost all studied exclusively pathologists specializing in genitourinary diseases and may not reflect general pathologists who may perform the majority of bladder cancer grading in the community. Patients initially diagnosed in the community often have their slides re-read at our institution prior to initiating treatment, since the management of non-muscle invasive bladder cancer relies primarily on tumor grade.

Due to concerns of persistent inter-observer variability using the contemporary grading system, we examined the reproducibility of both the 1973 and 2004 systems among pathologists at our institution including both specialists in genitourinary pathology as well as a general pathologist.

Materials and methods

Ninety-eight consecutive transurethral resection of bladder tumor (TURBT) specimens with non-invasive (Ta) disease were identified from the Department of Pathology at New York University Langone Medical Center (NYULMC) dating back to 2007, after approval by the institutional internal review board. All specimens originated from surgeries performed at NYULMC. The specimens had been previously embedded, sectioned, and stained with hematoxylin-eosin according to the departmental protocol, and one representative slide was selected from each case for review. Slides were de-identified and coded to ensure blinding of the pathologists. The slides were then read by two pathologists specializing in genitourinary pathology as well as by one general pathologist. The slide set was first read according to the 1973 WHO grading system, assigning tumors as papilloma, grade 1, grade 2, or grade 3. The entire set was then read according to the 2004 WHO/ISUP grading system, assigning tumors as papilloma, papillary urothelial neoplasm of low malignant potential (PUNLMP), low grade, and high grade. Slide review was performed at the discretion of the reviewing pathologist over the course of several sessions. Grade was assigned to reflect the least favorable grade encountered, which occupied at least 5% of the specimen, as per the convention of the Department of Pathology at NYULMC.

In addition to raw percentage agreement, the kappa statistic (κ) was used to measure inter-observer variability in order to account for agreement that would be expected purely by chance.[13] By convention, κ values from 0.21-0.40 represent "fair" agreement, from 0.41-0.60 represent "moderate" agreement, and > 0.60 represent "substantial agreement". (For reference, the κ value for prostate cancer Gleason grading was recently reported to be 0.76.[14] A generalized κ statistic and 95% confidence intervals were used to compare overall agreement of the three pathologists as well as agreement of each pairing using the 1973 system against the 2004 system. Microsoft Excel and SPSS were used to conduct statistical analyses.

Results

The distributions of grade assignments for the 98 specimens using the 1973 and 2004 systems are shown for each pathologist in Table 1. The raw percentage agreement of the three pathologists using the 1973 system was only 26% for the 98 specimens. When using the 2004 system, raw percentage agreement was still only 47%. There was no trend in terms of rate of agreement when comparing the first slides reviewed versus the last slides reviewed.

TABLE 1. **Distribution of 1973 grades per pathologist**

	GU1	GU2	GP
Papilloma	4	2	3
G1	41	19	61
G2	35	56	24
G3	18	21	10

GU1 = genitourinary specialist #1; GU2 = genitourinary specialist #2; GP = general pathologist

TABLE 2. **Distribution of 2004 grades per pathologist**

	GU1	GU2	GP
Papilloma	4	2	3
PUNLMP	7	6	3
Low grade	58	47	70
High grade	29	43	22

GU1 = genitourinary specialist #1; GU2 = genitourinary specialist #2; GP = general pathologist; PUNLMP = papillary urothelial neoplasm of low malignant potential

Table 2 shows the κ values for the whole group as well as for individual pairings, along with associated 95% confidence intervals. As evidenced by the overlapping confidence intervals, the 2004 system did not significantly improve overall agreement between all three pathologists, with κ = 0.26 for the 1973 system versus κ = 0.38 for the 2004 system. Both of these values represent only a "fair" level of agreement. There was also no significant improvement in agreement when calculated between the two specialists alone, with κ = 0.37 for the 1973 system and κ = 0.31 for the 2004 system. Again, both values represented only "fair" agreement. Agreement between the generalist with each specialist trended higher using the 2004 system over the 1973 system, but again this did not reach statistical significance (κ = 0.39 and κ = 0.46 using 2004, versus κ = 0.18 and κ = 0.33 using 1973), Table 3.

Discussion

Pathological grade and stage are relied upon most heavily in the determination of appropriate treatment for patients who present with UC. In the subset of non-invasive UC, which makes up about 70% of initial diagnoses, pathological grade is the single most important prognostic factor. The clinical implications of a low versus high grade designation are significant: aggressive pathological features warrant re-resection followed by intra-vesical immunotherapy with often toxic side effects, and consideration for early radical extirpative surgery in those with frequent recurrences. In contrast, less aggressive pathological features are reassuring to the clinician, who may then eschew the aforementioned treatments in favor of a conservative approach with periodic surveillance. Decisions regarding potentially invasive therapies are made based largely on pathological grade, and therefore it is of paramount importance that the pathologist is able to make an accurate and reproducible diagnosis of low versus high grade.

It is has been established that the WHO grading system of 1973 suffers from very poor inter-observer variability.[4] Moreover, it contains the grade 2 designation, which represents an intermediate level of aggressiveness between grade 1 and grade 3 and presents a management quandary to the clinician. The 2004 WHO/ISUP revision provides far more detailed pathological criteria for each grade, and also eliminates the intermediate grade 2 category by designating all lesions defined as carcinoma as either low or high grade.

Several recent reports have sought to determine whether these changes have made a significant improvement over the 1973 system in terms of reproducibility amongst pathologists. In 2003, Yorukoglu et al compared the agreement of six

TABLE 3. **Variability (κ) between pathologists using 1973 versus 2004 classifications**

	1973 κ (95% CI)	2004 κ (95% CI)
Overall	0.26 (0.16-0.39)	0.38 (0.25-0.50)
GU1 versus GU2	0.37 (0.22-0.51)	0.31 (0.16-0.46)
GU1 versus GP	0.33 (0.19-0.48)	0.46 (0.30-0.63)
GU2 versus GP	0.18 (0.07-0.31)	0.39 (0.26-0.54)

GP = general pathologist; GU1 = genitourinary specialist #1; GU2 = genitourinary specialist #2
κ values from 0.21-0.4 represent "fair", from 0.41-0.6 represent "moderate", and > 0.6 represent "substantial" agreement

urological pathologists grading 30 slides, comparing the 1973 system to what would become known as the 2004 revision.[11] They reported $\kappa = 0.56$ for the 2004 system, versus $\kappa = 0.48$ for the 1973, and concluded that the new grading system did not significantly improve reproducibility. Conversely, in 2009 May et al performed a larger, multi-center study with four urological pathologists grading 200 slides using both systems.[9] They did not calculate a generalized κ value for all four pathologists and instead listed κ values for all possible pairings of pathologists. They found κ values ranging from 0.003-0.365 for the 1973 system versus 0.296-0.516 for the 2004 system, and based upon this, they concluded that the 2004 system had less inter-observer variability. However, they did not report confidence intervals and therefore the slight differences seen in the κ value ranges for 1973 and 2004 are difficult to interpret. Nevertheless, the fact remains that all κ values were moderate at best ($\kappa < 0.6$) indicating that inter-observer variability remains an issue even with the 2004 system, which has been supported by other investigators.[12,15,16]

The present study suggests that the reproducibility of the 2004 revision not only failed to improve upon the 1973 system, but it also remains high for the determination of a prognostic factor used to dictate the management of superficial UC. Moreover, our study is the first to compare reproducibility between expert urological pathologists versus general pathologists, who may read the majority of bladder specimens outside of major academic centers. We found that there was no improvement in the concordance between pathologists with urological expertise as opposed to general pathologists, suggesting that the limitations in pathological grading of UC cannot be overcome with further training or experience. This is consistent with the findings of Murphy et al, who could not demonstrate a clear improvement after a period of intensive training among urological pathologists using the 2004 system.[17] Our findings also suggest that despite the revision, the 2004 system is not reproducible even in the most experienced hands.

In eliminating the intermediate grade 2 designation, the 2004 system became a 2-tier system for carcinoma (low versus high grade). Despite this, in the present study the three pathologists agreed in just 47% of the 98 cases. Therefore, more than half of the patients could receive a different pathological grade for their tumor depending on which pathologist grades their specimen. Particularly for superficial UC, the clinical implications cannot be understated, as a patient could undergo a second transurethral resection, induction BCG therapy, and even early cystectomy if the specimen is read as a high grade tumor versus interval cystoscopic management if it is read as low. For this reason, at NYULMC if a patient presents with a pathological diagnosis from an outside institution, our standard protocol is to have the slides re-read by our own pathologists prior to initiating treatment. The importance of tumor grade in dictating therapeutic decisions, as well as our data demonstrating poor grading reproducibility even among specialists, demands a more reliable means of identifying low and high risk tumors than a grading system based on morphology. An emerging understanding of the molecular basis for the divergent pathways of low and high grade UC may pave the way for a molecular grading system based on their distinctive genetic defects, with far greater prognostic ability.[3,8,18]

Conclusion

In conclusion, pathological grading remains at the cornerstone of risk stratification and subsequent treatment selection for superficial UC; however despite the 2004 revision, reproducibility among even expert urological pathologists is low. There is a great need to move beyond the "guessing game" of pathological grade based on morphological features, and to elucidate molecular markers based on the divergent biology of low versus high grade disease. Validation of such markers could potentially lead to a more reliable risk assessment tool and ultimately help guide more appropriate therapy of patients with UC. □

References

1. Howlader N, Noone AM, Krapcho M et al. SEER Cancer Statistics Review, 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2011/, based on November 2013 SEER data submission, posted to the SEER web site April 2014, accessed August 1, 2014.
2. Sylvester RJ, van der Meijden AP, Oosterlinck W et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* 2006; 49(3):466-477.
3. Wu XR. Urothelial tumorigenesis: a tale of divergent pathways. *Nat Rev Cancer* 2005;5(9):713-725.
4. Ooms ECM, Anderson WAD, Alons CL, Boon ME, Veldhizen RW. Analysis of the performance of pathologists in the grading of bladder tumors. *Hum Pathol* 1983;14(2):140-143.
5. Bostwick DG, Mikuz G. Urothelial papillary (exophytic) neoplasms. *Virchows Arch* 2002;441(2):109-116.
6. Epstein JI, Amin MB, Reuter VR, Mostofi FK. The World Heath Organization/International Society of Urological Pathology consensus classification of urothelial (transitional) cell neoplasms of the urinary bladder. Bladder Consensus Conference Committee. *Am J Surg Pathol* 1998;22(12):1435-1448.

7. Pellucchi F, Freschi M, Ibrahim B et al. Clinical reliability of the 2004 WHO histological classification system compared with the 1973 WHO system for Ta primary bladder tumors. *J Urol* 2011; 186(6):2194-2199.
8. Burger M, Van der Aa MNM, Van Oers JMM et al. Prediction of progression of non-muscle-invasive bladder cancer by WHO 1974 and 2004 grading and by FGFR3 mutation status: a prospective study. *Eur Urol* 2008;54(4):835-843.
9. May M, Brookman-Amissah S, Roigas J et al. Prognostic accuracy of individual uropathologists in noninvasive urinary bladder carcinoma: a multicentre study comparing the 1973 and 2004 World Health Organization classifications. *Eur Urol* 2010; 57(5):850-858.
10. van Rhijin BWG, van Leenders GJLH, Ooms BCM et al. The pathologist's mean grade is constant and individualizes the prognostic value of bladder cancer grading. *Eur Urol* 2010;57(6): 1052-1057.
11. Yorukoglu K, Tuna B, Dikicioglu E et al. Reproducibility of the 1998 World Health Organization/International Society of Urologic Pathology classification of papillary urothelial neoplasms of the urinary bladder. *Virchows Arch* 2003;443(6): 734-740.
12. Gonul II, Poyraz A, Unsal C, Acar C, Alkibay T. Comparison of 1998 WHO/ISUP and 1973 WHO classifications for interobserver variability in grading of papillary urothelial neoplasms of the bladder. *Urol Int* 2007;78(4):338-344.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-174.
14. McKenney JK, Simko J, Bonham M et al. The potential impact of reproducibility of Gleason grading in men with early stage prostate cancer managed by active surveillance: a multi-institutional study. *J Urol* 2011;186(2):465-469.
15. Bol MG, Baak JP, Buhr-Wildhagen S et al. Reproducibility and prognostic variability of grade and lamina propria invasion in stages Ta, T1 urothelial carcinoma of the bladder. *J Urol* 2003; 169(4):1291-1294.
16. Engers R. Reproducibility and reliability of tumor grading in urological neoplasms. *World J Urol* 2007;25(6):595-605.
17. Murphy WM, Takezawa K, Maruniak NA. Interobserver discrepancy using the 1998 World Health Organization/ International Society of Urologic Pathology classification of urothelial neoplasms: practical choices for patient care. *J Urol* 2002; 168(3):968-972.
18. Mitra AP, Datar RH, Cote RJ. Molecular pathways in invasive bladder cancer: new insights into mechanisms, progression, and target identification. *J Clin Onc* 2006; 24(35):5552-5564.