

Reproducibility and reliability of tumor grading in urological neoplasms

Rainer Engers

Received: 24 May 2007 / Accepted: 2 August 2007 / Published online: 9 September 2007
© Springer-Verlag 2007

Abstract Histopathologic tumor grading reflects the degree of differentiation of a given tumor and for most urological tumors grading is an important factor in predicting their biological aggressiveness. Consequently, the clinical management of tumor patients is often strongly influenced by the tumor grade, provided by pathologists. This implicates that an ideal grading system should not only be of high prognostic relevance, but also of high reproducibility among different pathologists. To this end individual histological grading systems have been developed for different tumor entities and even for a given tumor type several grading systems have been proposed. All of these grading systems possess an inherent degree of subjectivity and consequently, both intra- and interobserver variability exist. In this review, grading systems for the most frequent urological tumors (i.e. prostate cancer, renal cell carcinoma, and urothelial tumors) are mentioned and data on the reproducibility and reliability of the most commonly used grading systems are summarized.

Keywords Grading · Reproducibility · Reliability · Urological tumors · Prostate cancer · Renal cell carcinoma · Urothelial neoplasms

Introduction

Histopathological tumor grading reflects the degree of differentiation of a given tumor and for most solid tumors,

including urological tumors, grading is one of the most important prognostic markers. This is reflected by the fact that according to the working classification of prognostic factors, introduced by the College of American Pathologists [1], for many tumors, grading is classified as a category I prognostic factor, meaning that its prognostic value is well supported by the literature and that it is generally used in patient management. Prognostic factors of category II, instead, have extensively been studied biologically and/or clinically, but have not been conclusively proven to be of value in multivariate analyses. The remaining category III applies to those factors that show some promise but do not meet the criteria of categories I or II.

Given its strong prognostic impact in predicting the biological aggressiveness of malignant tumors, the tumor grade, provided by pathologists, strongly influences the clinical management of tumor patients. Consequently, an ideal grading system should meet at least two major requirements at the same time: it should be of high prognostic relevance and of high reproducibility among different pathologists. Therefore, several histopathological grading systems have been developed for different tumor entities as well as for a given tumor type. All of these grading systems have in common an inherent degree of subjectivity, resulting in both intra- and interobserver variability. The reproducibility of these grading systems among pathologists has been analysed in several studies. The comparability of these studies, however, is limited, as for example different statistical methods have been used. Whereas some studies only determined the percentage of agreement, others used the more elaborate kappa (κ) and weighted kappa (κ_w) analyses. Kappa (κ) and κ_w are very useful measures of interobserver agreement, as the level of agreement is adjusted for that expected by chance [2–4]. When the observed agreement exceeds chance agreement, κ is

R. Engers (✉)
Institute of Pathology,
University Hospital Duesseldorf, Moorenstr. 5,
40225 Duesseldorf, Germany
e-mail: engers@med.uni-duesseldorf.de

positive, with its magnitude reflecting the strength of agreement. Thus, κ 0.00–0.20 reflects slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement [2–4]. In addition to κ , κ_w uses weights to quantify the relative difference between categories. Close disagreement is not weighted as heavily as more serious disagreement.

In the present review, grading systems for the most frequent urological tumors (i.e. prostate cancer, renal cell carcinoma, and urothelial tumors) are mentioned and data on the reproducibility and reliability of the most commonly used grading systems are summarized.

Prostate cancer

In order to predict the clinical behaviour and aggressiveness of prostate cancer, several grading systems with proven prognostic relevance have been developed by pathologists in the past decades [5–10]. Among these, the Gleason grading system emerged to be universally acknowledged and most commonly used, and consequently it was also included in the new World Health Organization (WHO) classification [11]. Since its first description the Gleason grading system was slightly modified in the mid sixties [5, 12] and mid seventies of the last century [13, 14], and lastly in 2005 by the international society of urological pathology (ISUP) consensus conference on Gleason grading of prostate carcinoma [15].

To be of clinical relevance, a grading system must display sufficient reproducibility and grading done on biopsies should be reasonably representative for the tumor as a whole [7, 16]. The correlation between biopsy and prostatectomy Gleason scores has been investigated in several studies [17–20] and an exact concordance was only found in 28–68% (pooled data, 44.5%). This relatively poor concordance is mainly caused by undergrading of low-grade carcinomas in biopsy specimens, whereas higher agreement is achieved in grading of high-grade carcinomas [16, 21, 22]. More specifically, biopsies were found to be undergraded in 24–60% (pooled data 45%) and overgraded in 5–32% (pooled data, 10.4%) [17–20].

To improve this concordance, the Gleason grading system and its practical application by pathologists have recently been subjected to several modifications by an ISUP consensus conference [15]. Of these, the most important modifications refer to the definitions of Gleason pattern 3 and 4, respectively. Thus, “individual cells” are no longer allowed within Gleason pattern 3 and most cribriform patterns are diagnosed as Gleason pattern 4, with only rare cribriform lesions satisfying diagnostic criteria for cribriform pattern 3 (i.e. rounded, well-circumscribed glands of

the same size as normal glands). Ill-defined glands with poorly formed glandular lumina warrant the diagnosis of Gleason pattern 4, whereas very small well-formed glands still are within the spectrum of Gleason pattern 3. In addition, it was recommended that on needle biopsies with more than two different Gleason patterns, both the primary pattern (i.e. the most prevalent pattern) and the highest grade should be recorded. Furthermore, the diagnosis of Gleason score 4 on needle biopsies should only be made “rarely, if ever”. As a more detailed description of all modifications, introduced by the ISUP consensus conference, is beyond the scope of this review, for further information the reader is referred to the respective literature [15, 23]. Interestingly, a recent study showed that these modifications resulted in a significant improvement of concordance between biopsy and prostatectomy Gleason scores from 58 to 72%, when compared to conventional Gleason grading [16]. It is still unclear whether or not the overall concordance between biopsy and prostatectomy Gleason scores can also be significantly increased by using extended biopsy schemes. While an earlier study suggested that the prediction of the prostatectomy Gleason score is only marginally improved by increasing the number of biopsies [24], it was recently shown that the overall concordance between biopsy and prostatectomy Gleason scores can significantly be increased from 48 to 68% by using an extended biopsy scheme (mean 12.4 biopsy cores) rather than a traditional sextant scheme [25].

In addition to its representativity for the tumor as a whole, a clinically useful grading system must be sufficiently reproducible. The Gleason grading system, like all histological grading methods, possesses an inherent degree of subjectivity. Consequently, both intra- and interobserver variability exist. Intraobserver agreement of Gleason grades was reported in 43–78% of cases and agreement within ± 1 Gleason score unit was found in 72–87% of cases [26–28]. This is in line with the intraobserver agreement of Dr. Gleason himself, who wrote that on re-examining routine clinical material (not prototypical examples), including approximately 50% needle biopsies, he has duplicated exactly his previous histological scores approximately 50% of the time and within ± 1 histologic score (range 2–10) approximately 85% of the time [29]. In contrast to the limited number of studies on intraobserver agreement of Gleason grading, several studies addressed the question of interobserver reproducibility (Table 1). The comparability of these studies, however, is limited as the study designs vary partly considerably in terms of the definition of agreement (e.g. exact agreement, Gleason score ± 1 , or Gleason categories), the statistical analysis of agreement (e.g. percentage values, κ values, or weighted κ values), the type of specimens investigated (e.g. biopsies, radical prostatectomy specimens, transurethral resection specimens, a mixture of these, or tissue

Table 1 Literature review on interobserver agreement of Gleason grading

Study	No. of pathologists	Type of pathologists	Tissues	Agreement	κ values
Bain et al. [75]	7		9 biopsies 49 TUR	Exact: 47–60% ± 1 : 74–93%	± 1 : κ : 0.605–0.836
Svanholm et al. [76]	2		4 biopsies 87 TUR	Exact: 36% ± 1 : 69%	$\kappa = 0.467$
ten Kate et al. [77]	5	mixed	50 RP	3 groups: 2–4: 11% 5–7: 76% 8–10: 13%	κ_w : 0.30
Rousselet et al. [78]	2		234 specimens	Exact: 65% ± 1 : 94%	
de las Morenas et al. [79]	3		Biopsies or TUR	3 groups: 66%	
di Loreto et al. [80]	3		14 biopsies 12 TUR 15 RP	Exact: 46–56% ± 1 : 31–44%	
Ozdamar et al. [27]	2		96 specimens	Exact: 70.8%	
McLean et al. [81]	3		71 specimens	Exact: 43.7%	κ_w scores: 0.16–0.29 κ groups: 0.15–0.29
Lessells et al. [82]	12		46 biopsies		κ_w scores: 0.45 κ groups: 0.32
Egevad [34]	85	12 residents 73 specialists	40 images	Exact: 70.5%	$\kappa = 0.58$
Allsbrook et al. [31]	41	GP	38 biopsies		$\kappa = 0.435$ (0.00–0.88)
Allsbrook et al. [30]	10	UP	46 biopsies		κ_w scores: 0.56–0.70 κ groups: 0.47–0.64
Bova et al. [83]	2	Expert genitourinary pathologists	221 TMA images	Exact: 81% ± 1 : 100%	
de la Taille et al. [84]	10		104 TMA images	Exact: 21–82.5% 3 groups: 29–93%	κ : 0.28–0.54
	4		537 TMA images	Exact: 41–79% 3 groups: 52–88%	κ : 0.45–0.69
Mikami et al. [35]	61		16 specimens	Exact: 61.3% (12.5–100%)	0.44 (–0.47–1.0)
Oyama et al. [32]	14	8 GP 6 UP	37 biopsy images		GP (4 groups): 0.49 UP (4 groups): 0.68
Melia et al. [28]	9	UP	81 biopsies	4 groups: 78% (71.6–86.4%)	4 groups: 0.54 2 groups: 0.61

TUR transurethral resection, *RP* radical prostatectomy, *GP* general pathologists, *UP* urologic pathologists, *TMA* tissue microarray

microarray spots), the number of specimens investigated, the number of pathologists involved, and the qualification/specialisation of pathologists involved (e.g. general pathologists, genitourinary pathologists, or expert genitourinary pathologists). Overall the interobserver agreement in these studies was mostly moderate with κ values around 0.5–0.6 (Table 1). Given the high clinical relevance of Gleason grading for making treatment decisions, this level of agreement is unsatisfying. Apparently, the main reason for this situation is insufficient experience and familiarity of pathologists with this grading system. This is suggested by the fact that genitourinary pathologists or pathologists with special interests in genitourinary pathology seem to achieve higher

agreement levels with κ values of 0.6–0.8 than general pathologists [30–32]. In line with this, Allsbrook et al. found that pathologists, who learned Gleason grading at a meeting or a course, achieved higher agreements than pathologists who had not [31]. Overall, the major interobserver reproducibility problem is undergrading. In particular Gleason scores 5–6 are often underdiagnosed as Gleason scores 2–4, and cribriform sheets and fragments of Gleason pattern 4 are often mistaken for Gleason pattern 3 [30–32]. To overcome this problem different teaching methods have been developed and it was shown that this way the level of agreement in Gleason grading could be markedly improved from moderate to substantial with κ values ranging from 0.68 to 0.78

[33–35]. Moreover, a recent ISUP consensus conference developed new standards both in the definition of pattern characteristics and in the application of the Gleason grading system in general [15]. Application of these recommendations already led to increased concordance between biopsy and prostatectomy Gleason scores, when compared to conventional Gleason grading [16]. So far, however, only one study investigated the effects of these recommendations on interobserver agreement in Gleason grading. According to this study interobserver agreement of modified Gleason grading, as measured by κ w statistics in a cohort of 69 consecutive prostatectomy specimens, is at least as high as that of conventional Gleason grading with mean κ w being 0.58 and 0.56, respectively [36]. However, the effect of modified Gleason grading on intraobserver agreement still remains to be determined.

Renal cell carcinoma

Grading of renal cell carcinoma (RCC) was first reported in 1932 by Hand and Broders [37], describing that patients with high grade carcinoma were at higher risk to die than patients with low grade tumors. Since then, numerous studies have established the prognostic value of RCC grading and different grading systems emerged from these studies (for review see [38]). Of these, nuclear grading systems seem to be more predictive of disease-specific survival than grading systems based on cytoplasmic and/or architectural features of tumor cells [39–43]. As each system has its own advantages and disadvantages [38], there is no consensus yet as to which grading system should be used [42]. Rather in 1997 it has been stated by an International Consensus Conference that an ideal grading system, which preferentially should be a three-grade system, needs to be established [42]. By now, however, the four-tiered Fuhrman grading system is the most commonly used system in Europe and North America.

Studies on the reproducibility of RCC grading are limited and most of them refer to the Fuhrman grading system (Table 2). The first study addressing this topic was published by Lanigan et al. in 1994 [44]. In this study the authors compared the reproducibility of four different grading systems, including the Arner, Skinner, Syrjanen-Hjelt and Fuhrman systems, among four different pathologists [44]. Using the κ statistics of Landis and Koch [3] as a measure for interobserver agreement, which corrects for chance agreement, interobserver agreement was found to be fair to moderate with κ values of 0.42 (Syrjanen-Hjelt system), 0.33 (Fuhrman system), 0.26 (Skinner system) and 0.24 (Arner system), respectively. In a more recent study on 2042 RCCs, original nuclear grades as assigned at initial pathological diagnosis were compared to standardized

Table 2 Literature review on interobserver agreement of grading systems of renal cell carcinoma

Study	Grading system	No. of patients	Histological subtype	No. of pathologists	No. of grades	Agreement	κ values
Lanigan et al. [44]	Arner	88	any	4	3		0.24
	Skinner	88	any	4	4		0.26
	Syrjanen-Hjelt	88	any	4	4		0.42
	Fuhrman	88	any	4	4		0.33
Bretheau et al. [85]	Fuhrman	190	any	4	4	(concordance 95%)	
Lohse et al. [45]	Mayo Clinic standardized	1,733	clear cell RCC	>30 (original diagnosis) 1 (revision)	4	identical: 56.32% upshifted during revision: 35.26% downshifted during revision: 8.42%	
Al-Aynati et al. [49]	Fuhrman	99	any	4	4		0.29
					2 (G1-2 vs 3-4)		0.45
Ficarra et al. [47]	Fuhrman	388	any	5 (original diagnosis) 1 (revision)	4		0.44
Lang et al. [48]	Fuhrman	241	any	3	4		0.22 (0.09–0.36)
					3		0.26–0.34
					2 (G1-2 vs 3-4)		0.44

RCC renal cell carcinoma

nuclear grades reassigned after slide review and results were stratified for the different histological subtypes of RCC (i.e. clear cell, papillary and chromophobe cell type) [45]. For clear cell RCC nuclear grade remained unchanged on review for 56.32% of the tumors, while 35.26% were upshifted by one or more grades and 8.42% were downshifted on review. For papillary RCC nuclear grade remained unchanged for 49.1% of the tumors, whereas 44.1% were upshifted by one or more grades and 6.8% were downshifted. Similarly, for chromophobe cell RCC nuclear grade was unchanged in 55%, upshifted in 38% and downshifted in 7% of the tumors. κ values were not determined in this study. Of note, for all histological subtypes the reviewed grades were more predictive of death due to RCC than the respective original grades and this also held true after adjusting for the 1997 TNM stage. Recently, however, the relevance of Fuhrman grading in chromophobe cell RCC was questioned, as in a cohort of 87 cases Fuhrman grading failed to show any significant association with the patients' outcome [46].

In another study, original Fuhrman grades of 388 clear cell RCCs were compared to Fuhrman grades reassigned by a single pathologist after slide review [47]. Thus, tumors originally classified as G1 tumors were upshifted by 1 grade in 38.7% of the cases, by 2 grades in 18.9% of the cases and by 3 grades in 2.7% of the cases. Tumors originally classified as G2 were upshifted by 1 grade in 34% of the cases and by 2 grades in 4.3% of the cases. Grading of tumors originally classified as G3 and G4 remained unchanged in 73.1 and 89.3%, respectively. Overall, interobserver concordance in this study was moderate as indicated by a κ value of 0.44.

In a retrospective multicenter study, Lang et al. [48] investigated interobserver agreement of three pathologists in grading 241 RCCs according to the Fuhrman system. Using the original four-tiered Fuhrman grading system, a concordance rate among the three pathologists of only 24% was observed. The corresponding mean interobserver κ value was 0.22 (range 0.09–0.36), indicating fair agreement according to the commonly used interpretation of κ values, as proposed by Landis and Koch [3]. The level of concordance, however, could be improved by collapsing the original four-tiered grading system to three-tiered and two-tiered grading systems, respectively. Thus, the highest mean κ value was yielded by a two-tiered scheme in which low-grade (grade 1–2) tumors were distinguished from high-grade (grade 3–4) tumors. This way, agreement among all three pathologists occurred in 58.9% of the cases and a mean κ value of 0.44 (range 0.32–0.55) was achieved. Most importantly, the original 4-tiered Fuhrman grade was an independent prognostic factor for all three pathologists and nuclear grade continued to have independent prognostic value after the optimal collapsing algorithm was performed.

Similar results as reported by Lang et al. [48] were obtained in a study of Al-Aynati et al. [49]. In a cohort of 99 RCCs interobserver variability in four-tiered Fuhrman grading was determined among four pathologists and a mean κ value of 0.29 was observed. By combining Fuhrman grades 1 and 2 as low-grade tumors and grades 3 and 4 as high-grade tumors, interobserver agreement could be improved as indicated by a mean κ value of 0.45. In addition to their analysis on interobserver variability, the authors also addressed the question of intraobserver variability in the same study. To this end all four pathologists had to reassign Fuhrman grades to all cases after a period of 3–5 months. Using the four-tiered Fuhrman grading system, intraobserver κ values ranged from 0.29 to 0.62 (mean = 0.45), indicating a moderate level of concordance according to Landis and Koch [3]. When collapsing the diagnostic grades to 2 (low-grade tumors vs. high-grade tumors), intraobserver κ values ranged from 0.4 to 0.64 (mean = 0.53), reflecting a slight improvement within the range of moderate agreement.

Overall, interobserver agreement in nuclear grading of RCCs appears to be only fair to moderate, independent of the grading system used. This could at least in part be due to the heterogeneity of RCCs, as areas of different grades are often found within a given tumor [38, 49]. Nuclear grading should be based on the highest-grade area identified within a tumor. However, the minimum size required for such an area to be considered significant has not yet been standardized [38, 42, 49].

Urothelial neoplasms of the bladder

In the past decades, several classification and grading systems for urothelial neoplasms of the urinary bladder, including the 1973 WHO system [50], the Bergkvist system [51], the Murphy system [52], and the Pauwels system [53] have been proposed. Of these, the 1973 WHO classification and grading system has most commonly been used and remained unchanged for about 30 years. With regard to tumors, diagnosed as carcinomas, histologic grading was based on the degree of cellular anaplasia using a three-tiered scale: grade 1 (G1) was characterized by the least degree of anaplasia compatible with malignancy, grade 3 (G3) by the most severe degree of anaplasia, and grade 2 (G2) by an intermediate degree of anaplasia. The main problem with this grading system was the lack of defined cut-off points among the three grades of differentiation, giving rise to high intra- and interobserver variabilities (Table 3) [54–56]. This is reflected by the fact that the reported frequency of G2 tumors in non-selected tumor cohorts ranged from 13 to 69%, the frequency of G1 tumors from 8 to 25%, and the frequency of G3 tumors from 23 to

Table 3 Literature review on interobserver agreement of WHO grading systems of urothelial neoplasms

Study	Grading system	No. of cases	No. of pathologists	Agreement	κ values
Ooms et al. [59]	1973 WHO	57	6	0.46–0.58 (Spearman rank order classification)	
Abel et al. [86]	1973 WHO	99 re-graded cases	1	86.9%	
Robertson et al. [61]	1973 WHO	90	11		0.19–0.44
Tosoni et al. [60]	1973 WHO	301 re-graded cases	1	61%	
Murphy et al. [67]	1998 WHO/ISUP (=2004 WHO)	247	3		low-grade NIPUT: 0.12–0.5 high-grade urothelial tumors: 0.75–0.82
Campbell et al. [68]	1998 WHO/ISUP (=2004 WHO)	49 re-graded cases	1 2 (consensus)		$\kappa_w = 0.45$ $\kappa_w = 0.6$
Yorukoglu et al. [69]	1973 WHO vs. 1998 WHO/ISUP (=2004 WHO)	30	6		1973 WHO: 0.48 (0.19–0.65) 1998 WHO/ISUP (=2004 WHO): 0.56 (0.42–0.65)
Gönül et al. [56]	1973 WHO vs. 1998 WHO/ISUP (=2004 WHO)	258	2		<i>Overall:</i> 1973 WHO: 0.41 1998 WHO/ISUP (=2004 WHO): 0.59 <i>low-grade NIPUT:</i> 1973 WHO: 0.20 1998 WHO/ISUP (=2004 WHO): 0.26 <i>pT1 carcinomas:</i> 1973 WHO: 0.26 1998 WHO/ISUP (=2004 WHO): 0.91 <i>pT2 carcinomas:</i> 1973 WHO: no correlation 1998 WHO/ISUP (=2004 WHO): 0.31

WHO World Health Organization, ISUP International Society of Urological Pathology, NIPUT non-invasive papillary urothelial tumors

63% [53, 57, 58]. Ooms et al. [59] investigated interobserver variability in grading of bladder cancer among six pathologists in a setting of 57 cases. The Spearman rank-order correlations coefficients were 0.5–0.67 for intra- and 0.46–0.58 for interobserver variability and thus the results were interpreted by the authors as “disturbingly high” variability, which might invalidate the usefulness of grading in clinical decision making. Tosoni et al. [60] found interobserver discrepancies in grading in 39% of 301 cases of pTa and pT1 bladder cancers, respectively. In a study by Robertson et al. [61] interobserver agreement among 11 pathologists was slight to moderate as reflected by κ values ranging from 0.19 to 0.44.

Given the strong prognostic impact of tumor grading, this high variability raised concerns about the appropriateness of clinical management strategies in a setting of uncertainty about reliable tumor grade. Consequently, a new classification and grading system, subsequently known as the 1998 WHO/ISUP system, has been proposed in 1998 [54] and was adopted in 2004 in the most recent WHO classification and grading system (Pathology and genetics: tumours of the urinary system and male genital organs, one of a series of WHO “Blue Books”). The most important

changes in comparison to the 1973 WHO system were (1) the introduction of a new category of non-invasive papillary urothelial tumors, referred to as papillary urothelial neoplasms of low malignant potential (PUNLMP), (2) a detailed histological description of the different categories of non-invasive papillary urothelial tumors and (3) the collapsing of the formerly three-tiered grading system (G1, G2, G3) to a two-tiered system (low-grade vs. high-grade) for both non-invasive papillary carcinomas and invasive carcinomas in general. A comparison of the 1973 and 2004 WHO grading systems is shown in Fig. 1.

An important goal of the 2004 WHO classification was to improve reproducibility in diagnosis and grading among different pathologist by providing detailed histological criteria for each diagnostic category. To date a significant improvement in intra- and interobserver variability as compared to the 1973 WHO system has not been found (Table 3). Beyond others, this is reflected by the fact that in five different studies, in which non-invasive papillary urothelial tumors were graded according to the 1998 WHO/ISUP hence 2004 WHO system, the incidence of PUNLMP varied from 12 to 39%, that of low-grade carcinoma varied from 27 to 63%, and the incidence of high-grade carcinoma

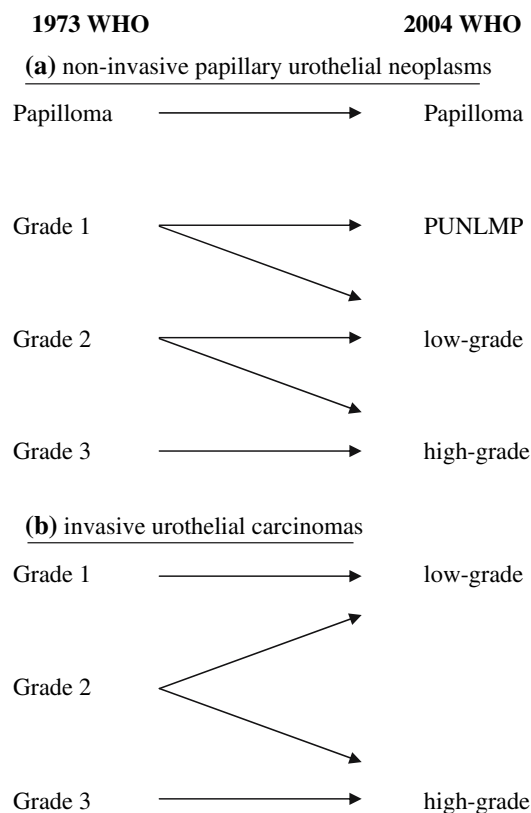


Fig. 1 Comparison of the 1973 and 2004 WHO grading system for non-invasive papillary urothelial neoplasms and invasive urothelial carcinomas in general. The 1973 WHO grade 1 non-invasive papillary urothelial carcinomas are reassigned, some to the papillary urothelial neoplasm of low malignant potential (PUNLMP) category, some to the low-grade non-invasive papillary urothelial carcinoma category. Grade 2 non-invasive papillary urothelial carcinomas (1973 WHO) are reassigned, some to the low-grade and the remaining to the high-grade non-invasive papillary urothelial carcinoma category. All grade 3 non-invasive papillary urothelial carcinomas (1973 WHO) are reassigned to the high-grade non-invasive papillary urothelial carcinomas. WHO World Health Organization, PUNLMP papillary urothelial neoplasm of low malignant potential

varied from 21 to 67% [62–66]. More specifically, Murphy et al. [67] found only slight to moderate interobserver agreement for PUNLMP and low-grade carcinomas among three pathologists ($\kappa = 0.12\text{--}0.50$), compared to substantial agreement for high-grade carcinomas and carcinoma in situ ($\kappa = 0.75\text{--}0.82$). In a study by Campbell et al. [68] interobserver variability of the 1998 WHO/ISUP system was found to be moderate ($\kappa = 0.45$) and the level of agreement could not significantly be increased even if the pathologists reviewed the cases together and reached a consensus diagnosis ($\kappa = 0.60$). Yorukoglu et al. [69] investigated intra- and interobserver agreement of both the 2004 WHO and the 1973 WHO system in a setting of 30 cases of non-invasive papillary urothelial tumors and six pathologists after having provided a teaching set to each study participant. No significant differences neither in intraobserver reproducibility

(2004 WHO: $\kappa = 0.67$, range 0.45–0.89 vs. 1973 WHO: $\kappa = 0.66$, range 0.45–0.89) nor in interobserver reproducibility (2004 WHO: $\kappa = 0.56$, range 0.42–0.65 vs. 1973 WHO: $\kappa = 0.48$, range 0.19–0.65) became evident. In a recent study of Gönül et al. [56] two pathologists assigned a tumor grade according to the 1973 WHO and the 1998 WHO/ISUP (=2004 WHO) system to 258 consecutive papillary urothelial carcinomas. Regardless of the pathologist, tumor grades of the two grading systems correlated to each other and to the pathological stage. The overall agreement between pathologists was somewhat higher in the 1998 WHO/ISUP (=2004 WHO) system ($\kappa = 0.59$) than in the 1973 WHO system ($\kappa = 0.41$), but both κ values were still within the range of a moderate agreement. Thus, in summary the studies performed so far suggest that the reproducibility of the 1998 WHO/ISUP (=2004 WHO) system does not appear to be appreciably different from that of the 1973 WHO classification.

Interestingly, however, in the study of Gönül et al. [56] the level of interobserver agreement of the 1998 WHO/ISUP (=2004 WHO) system considerably differed, when different tumor categories were compared. While the highest level of grading agreement was found in pT1 carcinomas ($\kappa = 0.91$), the lowest level of agreement was observed, when only tumors of the PUNLMP and the low-grade non-invasive papillary carcinoma categories were included in the analysis ($\kappa = 0.26$). Similarly, Murphy et al. [67] reported a 50% discrepancy rate among pathologists attempting to distinguish between PUNLMPs and low-grade papillary urothelial carcinomas even after a period of structured education. Accordingly, in a study of Yorukoglu et al. [69] mean rates of agreement for PUNLMP, low-grade non-invasive papillary urothelial carcinoma, and high-grade non-invasive papillary urothelial carcinoma were 48, 72.7, and 92%, respectively. Apparently, the yet unsatisfying overall levels of interobserver agreement in grading of non-invasive papillary urothelial carcinomas according to the 2004 WHO system can largely be attributed to the fact that the histologic distinction between PUNLMP and low-grade non-invasive papillary urothelial carcinoma causes major difficulties, even for experienced pathologists and although detailed histological criteria for these categories have been provided.

This raises the question as to whether a distinction between PUNLMP and low-grade non-invasive papillary urothelial carcinoma is of any prognostic and clinical use, because only such a relevance would justify to stick to this classification. Intriguingly, studies on the prognostic and clinical relevance of the 1998 WHO/ISUP (=2004 WHO) system were performed only after its publication. In these studies PUNLMPs have been reported to recur in up to 60% and to progress to invasive carcinoma in up to 8% of the cases, whereas low-grade non-invasive papillary urothelial

carcinomas recurred in up to 77% and progressed in up to 13% of the cases [55, 62, 68, 70–74]. Overall, differences in aggressiveness hence prognosis of PUNLMPs and low-grade non-invasive papillary urothelial carcinomas, reported so far, seem to be slight rather than pronounced. Consequently, the clinical management of patients with PUNLMP or low-grade non-invasive papillary urothelial carcinomas is currently similar if not identical [55]. From this one might conclude that no differentiation between these categories is needed. However, given the strong interobserver variability among pathologists in distinguishing between these two categories and given the knowledge about the biological heterogeneity of low-grade non-invasive papillary urothelial tumors (including PUNLMPs), it might be promising—prior to abandon this classification—to search for additional (e.g. molecular) markers, which together with the established histological criteria allow a more precise distinction between prognostic hence clinically relevant subgroups.

Another aspect contributing to interobserver variability in grading of urothelial tumors is the well known fact of tumor heterogeneity. Different grades are often found within a given tumor and in general the overall grade is based on the highest-grade area identified within a tumor. However, similar as for renal cell carcinoma, the minimum size required for such an area to be considered significant has not yet been standardized. Consequently, some pathologist assign a high tumor grade in any case that a high-grade area is present. In contrast other pathologists assign a high tumor grade only when the high-grade area comprises more than 5% of an otherwise low-grade tumor.

Conclusions

Like other tumors, urological tumors, are known to be both biologically and morphologically heterogeneous. Consequently, histological grading systems possess an inherent degree of subjectivity, giving rise to both intra- and interobserver variability. In general, reproducibility levels of the most commonly used grading systems of urological tumors are fair to moderate and grading of low-grade tumors provides more difficulties to pathologists than grading of high-grade tumors. Nevertheless for most urological tumors, it is well established that grading is an important factor in predicting their biological aggressiveness.

With regard to prostate cancer, structured education was shown to significantly improve reproducibility in Gleason grading and consequently several teaching facilities have recently been established. Fuhrman grading of renal cell carcinomas is only fairly to moderately reproducible and collapsing the original four-tiered grading system to a two-tiered grading system seems to improve the reproducibility

only marginally. Studies addressing the value of structured teaching in Fuhrman grading have not been reported yet. A more precise definition of how to grade heterogeneous tumors with special emphasis on the minimal amount of high-grade areas, required to upgrade an otherwise low-grade tumor, will help to improve grading reproducibility, but most likely only to a limited extent. Therefore, it appears that a new grading system, which possibly also includes molecular markers, needs to be established. Grading reproducibility of urothelial tumors using the 2004 WHO system appears to be largely hampered by the difficulty to distinguish between PUNLMP and low-grade non-invasive papillary urothelial carcinoma and studies suggest that this difficulty cannot be overcome by structured teaching. Apparently, the distinction between PUNLMP and low-grade non-invasive papillary urothelial carcinoma cannot reliably be made based on the so far established histological criteria and rather requires the identification of specific (e.g. molecular) markers. As long as no such markers are available and the prognostic hence clinical relevance of the distinction between PUNLMP and low-grade non-invasive papillary urothelial carcinoma has not been established, the new terminology used in the 2004 WHO classification is of questionable validity and utility.

References

1. Hammond ME, Fitzgibbons PL, Compton CC, Grignon DJ, Page DL, Fielding LP, Bostwick D, Pajak TF (2000) College of American pathologists conference XXXV: solid tumor prognostic factors—which, how and so what? Summary document and recommendations for implementation. Cancer committee and conference participants. *Arch Pathol Lab Med* 124:958–965
2. Cohen JA (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
3. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
4. Brennan P, Silman A (1992) Statistical methods for assessing observer variability in clinical measures. *BMJ* 304:1491–1494
5. Gleason DF (1966) Classification of prostatic carcinomas. *Cancer Chemother Rep* 50:125–128
6. Brawn PN, Ayala AG, Von Eschenbach AC, Hussey DH, Johnson DE (1982) Histologic grading study of prostate adenocarcinoma: the development of a new system and comparison with other methods—a preliminary study. *Cancer* 49:525–532
7. Bocking A, Kiehn J, Heinzl-Wach M (1982) Combined histologic grading of prostatic carcinoma. *Cancer* 50:288–294
8. Helpap B, Bocking A, Dhom G, Faul P, Kastendieck H, Leistenschneider W, Muller HA (1985) Classification, histological and cytological grading and assessment of regression grading in prostatic carcinomas. A recommendation of the pathologic-urological task force on prostatic carcinoma. *Pathologie* 6:3–7
9. Mostofi FK (1975) Grading of prostatic carcinoma. *Cancer Chemother Rep* 59:111–117
10. Mostofi FK, Sesterhenn IA, Sobin LH (1980) Histological typing of prostate tumours. In: International histological classification of tumours, No. 22. World Health Organization, Geneva

11. Epstein JI, Algaba F, Allsbrook WC Jr, Bastacky S, Boccon-Gibod L, De Marzo AM, Egevad L, Furosato M, Hamper UM, Helpap B, Humphrey PA, Iczkowski KA, Lopez-Beltran A, Montironi R, Rubin MA, Sakr WA, Samaratunga H, Parkin DM (2004) Acinar adenocarcinoma. In: Eble JN, Sauter G, Epstein JI, Sesterhenn IA (eds) World Health Organization classification of tumours. pathology and genetics: tumours of the urinary system and male genital organs. IARC, Lyon, France, pp 179–184
12. Mellinger GT, Gleason D, Bailar J III (1967) The histology and prognosis of prostatic cancer. *J Urol* 97:331–337
13. Gleason DF, Mellinger GT (1974) Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J Urol* 111:58–64
14. Mellinger GT (1977) Prognosis of prostatic carcinoma. *Recent Results Cancer Res* 61–72
15. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL (2005) The 2005 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. *Am J Surg Pathol* 29:1228–1242
16. Helpap B, Egevad L (2006) The significance of modified Gleason grading of prostatic carcinoma in biopsy and radical prostatectomy specimens. *Virchows Arch* 449:622–627
17. Carlson GD, Calvanese CB, Kahane H, Epstein JI (1998) Accuracy of biopsy Gleason scores from a large uropathology laboratory: use of a diagnostic protocol to minimize observer variability. *Urology* 51:525–529
18. Cookson MS, Fleshner NE, Soloway SM, Fair WR (1997) Correlation between Gleason score of needle biopsy and radical prostatectomy specimen: accuracy and clinical implications. *J Urol* 157:559–562
19. Spires SE, Cibull ML, Wood DP Jr, Miller S, Spires SM, Banks ER (1994) Gleason histologic grading in prostatic carcinoma. Correlation of 18-gauge core biopsy with prostatectomy. *Arch Pathol Lab Med* 118:705–708
20. Steinberg DM, Sauvageot J, Piantadosi S, Epstein JI (1997) Correlation of prostate needle biopsy and radical prostatectomy Gleason grade in academic and community settings. *Am J Surg Pathol* 21:566–576
21. Lopez-Beltran A, Mikuz G, Luque RJ, Mazzucchelli R, Montironi R (2006) Current practice of Gleason grading of prostate carcinoma. *Virchows Arch* 448:111–118
22. Montironi R, Mazzucchelli R, Scarpelli M, Lopez-Beltran A, Fellegara G, Algaba F (2005) Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. *BJU Int* 95:1146–1152
23. Egevad L, Allsbrook WC Jr, Epstein JI (2005) Current practice of Gleason grading among genitourinary pathologists. *Hum Pathol* 36:5–9
24. Egevad L, Norlen BJ, Norberg M (2001) The value of multiple core biopsies for predicting the Gleason score of prostate cancer. *BJU Int* 88:716–721
25. Mian BM, Lehr DJ, Moore CK, Fisher HA, Kaufman RP Jr, Ross JS, Jennings TA, Nazeer T (2006) Role of prostate biopsy schemes in accurate prediction of Gleason scores. *Urology* 67:379–383
26. Cintra ML, Billis A (1991) Histologic grading of prostatic adenocarcinoma: intraobserver reproducibility of the Mostofi, Gleason and Bocking grading systems. *Int Urol Nephrol* 23:449–454
27. Ozdamar SO, Sarikaya S, Yildiz L, Atilla MK, Kandemir B, Yildiz S (1996) Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. *Int Urol Nephrol* 28:73–77
28. Melia J, Moseley R, Ball RY, Griffiths DF, Grigor K, Harnden P, Jarmulowicz M, McWilliam LJ, Montironi R, Waller M, Moss S, Parkinson MC (2006) A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 48:644–654
29. Gleason DF (1992) Histologic grading of prostate cancer: a perspective. *Hum Pathol* 23:273–279
30. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, Bostwick DG, Humphrey PA, Jones EC, Reuter VE, Sakr W, Sesterhenn IA, Troncoso P, Wheeler TM, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol* 32:74–80
31. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol* 32:81–88
32. Oyama T, Allsbrook WC Jr, Kurokawa K, Matsuda H, Segawa A, Sano T, Suzuki K, Epstein JI (2005) A comparison of interobserver reproducibility of Gleason grading of prostatic carcinoma in Japan and the United States. *Arch Pathol Lab Med* 129:1004–1010
33. Kronz JD, Silberman MA, Allsbrook WC, Epstein JI (2000) A web-based tutorial improves practicing pathologists' Gleason grading of images of prostate carcinoma specimens obtained by needle biopsy: validation of a new medical education paradigm. *Cancer* 89:1818–1823
34. Egevad L (2001) Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images. *Urology* 57:291–295
35. Mikami Y, Manabe T, Epstein JI, Shiraiishi T, Furusato M, Tsuzuki T, Matsuno Y, Sasano H (2003) Accuracy of gleason grading by practicing pathologists and the impact of education on improving agreement. *Hum Pathol* 34:658–665
36. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L (2004) Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. *Virchows Arch* 445:17–21
37. Hand JR, Broders A (1932) Carcinoma of the kidney: the degree of malignancy in relation to factors bearing on prognosis. *J Urol* 28:199–216
38. Goldstein NS (1997) The current state of renal cell carcinoma grading. Union Internationale Contre le Cancer (UICC) and the American joint committee on cancer (AJCC). *Cancer* 80:977–980
39. Skinner DG, Colvin RB, Vermillion CD, Pfister RC, Leadbetter WF (1971) Diagnosis and management of renal cell carcinoma. A clinical and pathologic study of 309 cases. *Cancer* 28:1165–1177
40. Fuhrman SA, Lasky LC, Limas C (1982) Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol* 6:655–663
41. Thoenes W, Storkel S, Rumpelt HJ (1986) Histopathology and classification of renal cell tumors (adenomas, oncocytomas and carcinomas). The basic cytological and histopathological elements and their use for diagnostics. *Pathol Res Pract* 181:125–143
42. Medeiros LJ, Jones EC, Aizawa S, Aldape HC, Cheville JC, Goldstein NS, Lubensky IA, Ro J, Shanks J, Pacelli A, Jung SH (1997) Grading of renal cell carcinoma: workgroup No. 2. Union Internationale Contre le Cancer and the American Joint Committee on Cancer (AJCC). *Cancer* 80:990–991
43. Storkel S, Thoenes W, Jacobi GH, Lippold R (1989) Prognostic parameters in renal cell carcinoma—a new approach. *Eur Urol* 16:416–422
44. Lanigan D, Conroy R, Barry-Walsh C, Loftus B, Royston D, Leader M (1994) A comparative analysis of grading systems in renal adenocarcinoma. *Histopathology* 24:473–476
45. Lohse CM, Blute ML, Zincke H, Weaver AL, Cheville JC (2002) Comparison of standardized and nonstandardized nuclear grade of renal cell carcinoma to predict outcome among 2,042 patients. *Am J Clin Pathol* 118:877–886
46. Delahunt B, Sika-Paotonu D, Bethwaite PB, McCredie MR, Martignoni G, Eble JN, Jordan TW (2007) Fuhrman grading is not

- appropriate for chromophobe renal cell carcinoma. *Am J Surg Pathol* 31:957–960
47. Ficarra V, Martignoni G, Maffei N, Brunelli M, Novara G, Zanolla L, Pea M, Artibani W (2005) Original and reviewed nuclear grading according to the Fuhrman system: a multivariate analysis of 388 patients with conventional renal cell carcinoma. *Cancer* 103:68–75
 48. Lang H, Lindner V, de Fromont M, Molinie V, Letourneux H, Meyer N, Martin M, Jacqmin D (2005) Multicenter determination of optimal interobserver agreement using the Fuhrman grading system for renal cell carcinoma: assessment of 241 patients with >15-year follow-up. *Cancer* 103:625–629
 49. Al Aynati M, Chen V, Salama S, Shuhaibar H, Treleaven D, Vincic L (2003) Interobserver and intraobserver variability using the Fuhrman grading system for renal cell carcinoma. *Arch Pathol Lab Med* 127:593–596
 50. Mostofi FK, Sobin LH, Torloni H (1973) Histological typing of urinary bladder tumors, vol 10. World Health Organization, Geneva
 51. Bergkvist A, Ljungqvist A, Moberger G (1965) Classification of bladder tumours based on the cellular pattern. Preliminary report of a clinical-pathological study of 300 cases with a minimum follow-up of eight years. *Acta Chir Scand* 130:371–378
 52. Murphy WM (1989) Diseases of the urinary bladder, urethra, ureters, and renal pelvis. In: Murphy WM (ed) *Urological pathology*. WB Saunders, Philadelphia, pp 64–96
 53. Pauwels RP, Schapers RF, Smeets AW, Debruyne FM, Geraedts JP (1988) Grading in superficial bladder cancer. (1). Morphological criteria. *Br J Urol* 61:129–134
 54. Epstein JI, Amin MB, Reuter VR, Mostofi FK (1998) The World Health 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* 22:1435–1448
 55. MacLennan GT, Kirkali Z, Cheng L (2007) Histologic grading of noninvasive papillary urothelial neoplasms. *Eur Urol* 51:889–897
 56. Gönül II, Poyraz A, Ünsal C, Acar C, Alkibay T (2007) Comparison of 1998 WHO/ISUP and 1973 WHO classifications for interobserver variability in grading of papillary urothelial neoplasms of the bladder. *Pathological evaluation of 258 Cases*. *Urol Int* 78:338–344
 57. Jordan AM, Weingarten J, Murphy WM (1987) Transitional cell neoplasms of the urinary bladder. Can biologic potential be predicted from histologic grading? *Cancer* 60:2766–2774
 58. Lipponen PK (1992) Histological and quantitative prognostic factors in transitional cell bladder cancer treated by cystectomy. *Anticancer Res* 12:1527–1532
 59. Ooms EC, Anderson WA, Alons CL, Boon ME, Veldhuizen RW (1983) Analysis of the performance of pathologists in the grading of bladder tumors. *Hum Pathol* 14:140–143
 60. Tosoni I, Wagner U, Sauter G, Egloff M, Knonagel H, Alund G, Bannwart F, Mihatsch MJ, Gasser TC, Maurer R (2000) Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. *BJU Int* 85:48–53
 61. Robertson AJ, Beck JS, Burnett RA, Howatson SR, Lee FD, Lessells AM, McLaren KM, Moss SM, Simpson JG, Smith GD (1990) Observer variability in histopathological reporting of transitional cell carcinoma and epithelial dysplasia in bladders. *J Clin Pathol* 43:17–21
 62. Samaratunga H, Makarov DV, Epstein JI (2002) Comparison of WHO/ISUP and WHO classification of noninvasive papillary urothelial neoplasms for risk of progression. *Urology* 60:315–319
 63. Whisnant RE, Bastacky SI, Ohori NP (2003) Cytologic diagnosis of low-grade papillary urothelial neoplasms (low malignant potential and low-grade carcinoma) in the context of the 1998 WHO/ISUP classification. *Diagn Cytopathol* 28:186–190
 64. Bircan S, Candir O, Serel TA (2004) Comparison of WHO 1973, WHO/ISUP 1998, WHO 1999 grade and combined scoring systems in evaluation of bladder carcinoma. *Urol Int* 73:201–208
 65. Yin H, Leong AS (2004) Histologic grading of noninvasive papillary urothelial tumors: validation of the 1998 WHO/ISUP system by immunophenotyping and follow-up. *Am J Clin Pathol* 121:679–687
 66. Curry JL, Wojcik EM (2002) The effects of the current World Health Organization/International Society of Urologic Pathologists bladder neoplasm classification system on urine cytology results. *Cancer* 96:140–145
 67. Murphy WM, Takezawa K, Maruniak NA (2002) Interobserver discrepancy using the 1998 World Health Organization/International Society of Urologic Pathology classification of urothelial neoplasms: practical choices for patient care. *J Urol* 168:968–972
 68. Campbell PA, Conrad RJ, Campbell CM, Nicol DL, MacTaggart P (2004) Papillary urothelial neoplasm of low malignant potential: reliability of diagnosis and outcome. *BJU Int* 93:1228–1231
 69. Yorukoglu K, Tuna B, Dikicioglu E, Duzcan E, Isisag A, Sen S, Mungan U, Kirkali Z (2003) Reproducibility of the 1998 World Health Organization/International Society of Urologic Pathology classification of papillary urothelial neoplasms of the urinary bladder. *Virchows Arch* 443:734–740
 70. Cheng L, Neumann RM, Bostwick DG (1999) Papillary urothelial neoplasms of low malignant potential. Clinical and biologic implications. *Cancer* 86:2102–2108
 71. Holmang S, Andius P, Hedelin H, Wester K, Busch C, Johansson SL (2001) Stage progression in Ta papillary urothelial tumors: relationship to grade, immunohistochemical expression of tumor markers, mitotic frequency and DNA ploidy. *J Urol* 165:1124–1128
 72. Holmang S, Hedelin H, Anderstrom C, Holmberg E, Busch C, Johansson SL (1999) Recurrence and progression in low grade papillary urothelial tumors. *J Urol* 162:702–707
 73. Pich A, Chiusa L, Formiconi A, Galliano D, Bortolin P, Comino A, Navone R (2002) Proliferative activity is the most significant predictor of recurrence in noninvasive papillary urothelial neoplasms of low malignant potential and grade I papillary carcinomas of the bladder. *Cancer* 95:784–790
 74. Fujii Y, Kawakami S, Koga F, Nemoto T, Kihara K (2003) Long-term outcome of bladder papillary urothelial neoplasms of low malignant potential. *BJU Int* 92:559–562
 75. Bain GO, Koch M, Hanson J (1982) Feasibility of grading carcinomas. *Arch Pathol Lab Med* 106:265–267
 76. Svanholm H, Mygind H (1985) Prostatic carcinoma reproducibility of histologic grading. *Acta Pathol Microbiol Immunol Scand [A]* 93:67–71
 77. ten Kate FJW, Gallee MPW, Schmitz PIM, Joebis AC, van der Heul RO, Prins MEF, Blom JHM (1986) Problems in grading of prostatic carcinoma: interobserver reproducibility of five different grading systems. *World J Urol* 4:147–152
 78. Rousselet MC, Saint-Andre JP, Six P, Soret JY (1986) Reproducibility and prognostic value of Gleason's and Gaeta's histological grades in prostatic carcinoma. *Ann Urol (Paris)* 20:317–322
 79. de las Morenas A, Siroky MB, Merriam J, Stilmant MM (1988) Prostatic adenocarcinoma: reproducibility and correlation with clinical stages of four grading systems. *Hum Pathol* 19:595–597
 80. di Loreto C, Fitzpatrick B, Underhill S, Kim DH, Dytch HE, Galera-Davidson H, Bibbo M (1991) Correlation between visual clues, objective architectural features, and interobserver agreement in prostate cancer. *Am J Clin Pathol* 96:70–75
 81. McLean M, Srigley J, Banerjee D, Warde P, Hao Y (1997) Interobserver variation in prostate cancer Gleason scoring: are there implications for the design of clinical trials and treatment strategies? *Clin Oncol (R Coll Radiol)* 9:222–225

82. Lessells AM, Burnett RA, Howatson SR, Lang S, Lee FD, McLaren KM, Nairn ER, Ogston SA, Robertson AJ, Simpson JG, Smith GD, Tavadia HB, Walker F (1997) Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. *Hum Pathol* 28:646–649
83. Bova GS, Parmigiani G, Epstein JI, Wheeler T, Mucci NR, Rubin MA (2001) Web-based tissue microarray image data analysis: initial validation testing through prostate cancer Gleason grading. *Hum Pathol* 32:417–427
84. De La TA, Viellefond A, Berger N, Boucher E, De Fromont M, Fondimare A, Molinie V, Piron D, Sibony M, Staroz F, Triller M, Peltier E, Thiounn N, Rubin MA (2003) Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays. *Hum Pathol* 34:444–449
85. Bretheau D, Lechevallier E, de Fromont M, Sault MC, Rampal M, Coulange C (1995) Prognostic value of nuclear grade of renal cell carcinoma. *Cancer* 76:2543–2549
86. Abel PD, Henderson D, Bennett MK, Hall RR, Williams G (1988) Differing interpretations by pathologists of the pT category and grade of transitional cell cancer of the bladder. *Br J Urol* 62:339–342