Interobserver variability in Gleason histological grading of prostate cancer

Tayyar A. Ozkan, Ahmet T. Eruyar, Oguz O. Cebeci, Omur Memik, Levent Ozcan & Ibrahim Kuskonmaz

To cite this article: Tayyar A. Ozkan, Ahmet T. Eruyar, Oguz O. Cebeci, Omur Memik, Levent Ozcan & Ibrahim Kuskonmaz (2016): Interobserver variability in Gleason histological grading of prostate cancer, Scandinavian Journal of Urology, DOI: 10.1080/21681805.2016.1206619

To link to this article: http://dx.doi.org/10.1080/21681805.2016.1206619

Published online: 14 Jul 2016.

Submit your article to this journal

Article views: 3

View related articles

View Crossmark data
Interobserver variability in Gleason histological grading of prostate cancer

Tayyar A. Ozkan, Ahmet T. Eruyar, Oguz O. Cebeci, Omur Memika, Levent Ozcan and Ibrahim Kuskonmaz

ABSTRACT

Objective: The aims of this study were to evaluate the reproducibility of the Gleason grading system and to compare its interobserver variability with the novel Gleason grade grouping proposal using a large sample volume.

Materials and methods: In total, 407 pathology slides of prostate needle biopsies from 34 consecutive patients with prostate cancer were re-evaluated. The International Society of Urological Pathology 2005 modified Gleason grading system with Epstein’s modification was used. Two pathologists, blind to each other and to the initial pathology report, performed the pathological evaluation. To determine interobserver concordance, the kappa (κ) coefficient test was used.

Results: Pathologist 1 and pathologist 2 detected a tumor in 202 and 231 cores, respectively (p < 0.001). The two pathologists disagreed on the presence of a tumor in 31 cores. Of these 31 cores, 74% (n = 23/31) were Gleason pattern 3. The mean length of the cancer foci in these 31 disputed cores was 1.54 ± 0.8 mm. Concordance rates between the two observers for primary and secondary Gleason patterns were 63.96% (κ = 0.34) and 63.45% (κ = 0.37), respectively. Concordance with respect to the Gleason sum was 57.9% (κ = 0.43). When the Gleason scores were classified into the novel Gleason grade grouping, concordance was found to be 51.7% (κ = 0.39).

Conclusions: The agreement between observers on the Gleason sum was moderate. The novel Gleason grade grouping did not improve interobserver agreement. Further studies are needed to confirm these results on interobserver variability.

Introduction

Prostate adenocarcinoma [prostate cancer (PCa)] is the most common male cancer and the second leading cause of death in men. Diagnosis and grading of these heterogeneous tumors, which may morphologically form various cell structures, are very important in determining the biological behavior of the tumor and treatment modalities, using various nomograms. The Gleason histological grading system was developed based on the structural features formed by prostate tumor cells. It is the most commonly used histological grading system and is recommended by the World Health Organization. Gleason grading is an independent variable for predicting prognosis and determining the most suitable management for PCa treatment.

Histological grading systems should be easy to use and repeatable, so that the usefulness of clinically important prognostic information may be ensured [1]. A problem common to all grading systems is variability related to interobserver and intraobserver factors. As a result, many studies have been performed to evaluate interobserver concordance levels [2–8]. To enhance risk stratification and to enable a more accurate diagnosis on particular histological patterns, Gleason grade grouping according to prognostic significance has been proposed [9,10]. Gleason grading has been updated in accordance with the advice in the 2014 International Society of Urological Pathology (ISUP) Consensus Conference [11]. As proposed by the ISUP consensus report, in a recent study the Gleason patterns were classified into Gleason grade grouping (GGG) as follows: Gleason score 6 as GGG 1, Gleason score 3 + 4 = 7 as GGG 2, Gleason score 4 + 3 = 7 as GGG 3, Gleason score 8 as GGG 4, and Gleason score 9–10 as GGG 5 [12].

The current study aimed to evaluate the reproducibility of the Gleason grading system between two pathologists. To the authors’ knowledge, this is the first report to compare interobserver variability for the novel proposed Gleason grade grouping with a large volume of samples.

Materials and methods

The patients included in the study underwent 12-core prostate needle biopsies, and were treated with either prostatectomy (n = 18) or external beam radiotherapy (n = 13), or were currently in active surveillance (n = 3). Pathology slides of prostate needle biopsies from 34 consecutive patients with PCa were re-evaluated. A third investigator renumbered all the slides to blind both observers from descriptive clinical and identifying information. Each slide was...
re-marked with two numbers: one identifying the patient and the other identifying the core biopsy. Each pathology slide was assessed for both Gleason grades and scores on biopsy cores. Reproducibility was measured for both Gleason grades and scores individually. One slide of a single core, which had been severely damaged owing to an archiving error, was excluded. In total, 407 slides were re-evaluated.

Two pathologists who had been trained in different clinics were enrolled in the study. Both pathologists were trained with a webcast lecture covering the ISUP 2005 modified Gleason histological grading system and Epstein’s modification. According to Epstein’s proposal, all cribriform patterns were classified as Gleason pattern 4 rather than pattern 3, and glomeruloid glands were accepted as an early phase of cribriform pattern cancer and graded as pattern 4 [13]. The ISUP 2005 Gleason histological grading system with Epstein’s modification was used for pathological evaluation [14]. These two pathologists, blind to each other and the initial pathology report, performed pathological evaluation of the slides.

### Results

In total, 407 slides were evaluated. Pathologist 1 (Pat1) and pathologist 2 (Pat2) detected a tumor in 202 and 231 cores, respectively ($p < 0.001$). Both observers agreed on tumor presence in 197 cores. Pat1 and Pat2 disagreed on the presence of a tumor in 31 cores. Of these 31 cores, 74% ($n = 23/31$) were Gleason pattern 3. The mean length of the cancer foci in these 31 disputed cores was $1.54 \pm 0.8$ mm.

Neither pathologist detected Gleason pattern 1 or 2 tumors in any core. The most commonly reported Gleason sums (primary + secondary pattern) for Pat1 and Pat2 were 7 ($n = 85/202$) and 8 ($n = 64/231$), respectively ($p = 0.03$).

Pat1 recorded 25 cores as Gleason score $\leq 3 + 4$, whereas Pat2 recorded these cores as Gleason score $= 4 + 3$ ($n = 9$) or higher ($n = 16$). Three of 16 cores recorded as Gleason score $\geq 4 + 4$ measured cancer length were less than 1 mm and contained focal Gleason pattern 5 tumors (Figure 1(A–D)). Pat2 recorded 27 cores as Gleason score $\leq 3 + 4$; Pat1 recorded these same cores as Gleason score $= 4 + 3$ ($n = 18$) or higher grade ($n = 9$).

Concordance rates between the two observers for primary and secondary Gleason patterns were 64% ($\kappa = 0.34$, 95% CI 0.28–0.54) and 63% ($\kappa = 0.37$, 95% CI 0.31–0.42), respectively. Concordance with respect to the Gleason sum was 58% ($\kappa = 0.43$, 95% CI 0.42–0.48) (Figure 2(A)), and best concordance was at Gleason sum 7 (45.9%). When the Gleason scores were classified into Gleason grade grouping, concordance was found to be 52% ($\kappa = 0.39$, 95% CI 0.34–0.47) (Figure 2(B)). For novel Gleason grade grouping the best concordance was at GGG 4, which stands for Gleason score 4 + 4 (36%). The best concordance at Gleason sum 7 was distributed among four cells as in a 2 by 2 chart in Figure 2(B) for novel Gleason grade grouping. For this reason, a high concordance was no longer seen for either GGG 2 (Gleason score 3 + 4) or GGG 3 (Gleason score 4 + 3).

![Figure 1](image_url)  
Figure 1. (A) Prostate needle biopsy of prostate adenocarcinoma with different Gleason patterns (40×, H&E); (B) Gleason pattern 3 (200×, H&E); (C) Gleason pattern 4 (200×, H&E); (D) Gleason pattern 5 (200×, H&E).
Discussion

Histopathological grading systems are subjective methods owing to intraobserver and interobserver variability. Accurate diagnosis has been identified as one of the most important problems in previous interobserver agreement studies [1]. Three different methods have been defined for accurate diagnosis in such studies: calculating a complete interobserver agreement as a percentage, accepting the findings of an experienced observer as accurate, and consensus diagnosis. In the current study, interobserver agreement was evaluated using a kappa test, calculated as a percentage, and the results were accepted as an accurate diagnosis.

Figure 2. (A) Gleason sum assignments of the pathologists; (B) Gleason grade grouping assignments of the pathologists. Displayed numerals represent the percentage concordance rate (n) between the pathologists.
In the current study, pathologists reviewed all slides whether or not a tumor presented. Evaluation of these 407 slides resulted in inconsistent grades with respect to tumor presence in 31 cores. Nevertheless, there was moderate concordance with respect to the Gleason sum, as might be expected in a situation that parallels actual practice, where there is a high volume of slides, only some of which have tumors.

In a previous study, Mikami et al. used a tutorial atlas to determine the influence of education on reproducibility of the Gleason grading system. Sixteen slides were reviewed by two groups of pathologists, one of whom had had a Gleason grading tutorial, while the other had not. Improved interobserver variability was reported for the group that completed the tutorial on Gleason grading [5]. In the current study, both pathologists viewed a webcast lecture on the ISUP 2005 modified Gleason histological grading system and Epstein’s modification. The authors did not have an opportunity to assess the effect of an educational program, but according to the observers’ statements, the lecture improved their reports and concordance, particularly on low-grade cores.

In a multinational study, Oyama et al. compared interobserver reproducibility of the Gleason histological grading system. Thirty-seven identical images of pathological slides were digitally shared with six uropathologists and eight general pathologists. Gleason patterns were categorized into four groups (Gleason pattern 2–4, 5–6, 7, and 8–10). Four-tiered grading reported kappa coefficients for uropathologists and general pathologists as 0.68 and 0.49, respectively. In the same study, the authors concluded that cribriform sheets and fragments of cribriform Gleason pattern 4 carcinomas were undergraded as Gleason pattern 3 carcinomas [6]. This was discussed in another study; it is also a common problem among experienced uropathologists. Allsbrook et al. concluded that more precise descriptions should be made for low Gleason pattern tumors [2]. A multinational study reported on this particular problem of grading Gleason pattern 3 and 4 PCa with cribriform, fused or poorly formed glands. Fifteen experts in urological oncological pathology from 11 countries reviewed 25 slides of Gleason pattern 6–7 tumors. Gleason score results were grouped as Gleason pattern 5–6, 7 (Gleason score 3 + 4), 7 (Gleason score 4 + 3) and 8–10. The reported concordance among experts was moderate (κ = 0.43) [15]. Grading starts with detection and assessment of the tumor according to predefined histopathological characteristics. Although grade characteristics were previously defined and were well known by the observers, they are still subjective. Even in a relatively controlled study environment, agreement between experienced uropathologists is variable (κ = 0.43–0.68). On this particular issue, Helpap et al. [16] defined a novel gland fusion criterion as “no stromal connective tissue strands or bridges and only a single line of nuclei or remnants thereof and hence no traceable virtual line between at least two glands, at least two distinct and diverse gland lumina”. They demonstrated high concordance for both interobserver reproducibility on gland fusion and Gleason score 3 + 3 = 6 and 3 + 4 = 7 for microfocal PCa specimens between uropathologists. Overall, for both uropathologists and general pathologists, interobserver reproducibility for gland fusion was increased from 0.48 to 0.61 with increasing magnification (50×, 100× and 200×) [16]. In the current study, moderate interobserver agreement was found between pathologists for Gleason scores (κ = 0.43). The expected agreement between observers may be lower than the results of this study in real life. The most common problem encountered by the pathologists was their inability to differentiate between invasive growth and benign ductus/acinar structures, which were interpreted as Gleason pattern 3. Future studies on the Gleason grading system should consider a better definition for the grading of small foci. Epstein proposed that Gleason scores 2–4 should not be assigned to PCa on needle biopsy specimens, based on his observation of undergrading of higher grade carcinomas and variability among observers. Undergrading high-grade carcinomas may have an adverse impact on patient care [17]. Since that time, Gleason score 2–4 has been abandoned in needle biopsy reports by many pathologists [1–3,6–8,18].

Besides low-grade tumors, grading Gleason pattern 4 and 5 tumors presents some other difficulties. Similarly to previous studies, in the present study concordance on Gleason pattern 4 and 5 tumors was very low (18%) [1,3]. Indefinite acinar structures, cords, nesting of high-grade tumors and indeterminate borders of the tumor may be possible explanations for this low concordance.

In the reported studies, Gleason score concordance ranged from 9.9% to 70.8% [1,3,18–21]. In the current study, concordance for primary and secondary Gleason patterns had only three categories (Gleason patterns 3, 4 and 5). Concordance for both Gleason primary and secondary patterns was above 60%, although kappa coefficients were 0.34 and 0.37, respectively. For the Gleason sums, concordance was lower than primary and secondary patterns separately; however, a better kappa coefficient (κ = 0.43) was noted. In statistical analysis of observer variability, the kappa coefficient corrects chance. Fewer categories increase the concordance, but the kappa coefficient decreases, as shown in the present results. Including the current study, the reported studies used variable methodologies and grouping Gleason grades, which caused a wide range of interobserver agreement [6]. Epstein et al. recently proposed a novel Gleason grading system based on simplified grouping. It was concluded that this proposed Gleason grade grouping approach helps in obtaining more accurate grade stratification for outcomes and is simpler than the current system [12]. Loeb et al. have confirmed the predictive accuracy of the novel Gleason grade grouping in a large nationwide cohort [22]. Both studies concluded that the novel Gleason grade grouping has similar predictive accuracy and it was suggested as a simplified and user-friendly classification for patient counseling. In the current study, concordance using the novel five-tier Gleason grade grouping was 51.7% (κ = 0.39), although the Gleason sum kappa (κ = 0.43) was higher than the novel Gleason grade grouping. The proposed Gleason grade grouping did not yield a benefit for better interobserver concordance in this study. The debate over Gleason grading with respect to Gleason pattern 3 and 4 is still ongoing. Subdividing Gleason sum 7 into GGG 2 (Gleason score 3 + 4) and GGG 3 (Gleason score 4 + 3) resulted in
lower interobserver agreement for the proposed Gleason grade grouping.

Limitations of the study are the number of observers, the absence of experienced uropathologists, and unconfirmed histopathological results for external beam radiotherapy and active surveillance groups.

The most important problems in the evaluation of biopsy specimens are transitional areas between carcinoma foci, differentiation of benign ductus/acinar structures mimicking Gleason pattern 3, grading of low-grade small foci in fragmented material and changes related to compression artifacts, which are quite common in needle biopsies. The agreement between observers on the Gleason sum was moderate. The novel Gleason grade grouping and description, especially on low-grade tumors, did not improve interobserver agreement.

Acknowledgements

We wish to thank Murat Sener MD, Bekir Voyvoda MD and Emre Ulukaradag MD for their invaluable help and support for the study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

ORCID

Tayyar A. Ozkan http://orcid.org/0000-0003-3115-3141

References