North Clin Istanb 2025;12(3):337–343
doi: 10.14744/nci.2025.11456

Original Article PATHOLOGY

Assessment of interobserver variability in Gleason grading for prostate carcinoma

Ilkay Cinar, Esma Cinar

Department of Pathology, Giresun University Faculty of Medicine, Giresun, Turkiye

ABSTRACT

OBJECTIVE: The Gleason Score is the most widely used grading system for prostate adenocarcinoma and it is the strongest predictor of the patient's clinical outcome similar to other grading systems, and plays a key role in determining the most effective treatment strategy for the patient. The Gleason scoring system is subject to both intraobserver and interobserver variability. This study aims to assess the interobserver agreement for prostate adenocarcinoma within the Gleason grading system at our center, as well as identify contributing factors.

METHODS: A total of 119 cases diagnosed with prostatic adenocarcinoma at Giresun Training and Research Hospital were included in the study. Tissue samples from the cases had been subjected to routine laboratory procedures; three-micron sections were obtained from formalin-fixed paraffin blocks and stained H&E. Statistical investigation was conducted on the agreement between Gleason pattern, Gleason sum score, and grade group data among three observers.

RESULTS: In the evaluation, interobserver agreement was found to be minimal (Gleason pattern k=0.285, total Gleason sum score k=0.309, Grade group k=0.313). The assessment indicated higher agreement in determining low grade compared to high grade, with a decrease in interobserver agreement as the grade increased. Moreover, interobserver agreement demonstrated an increase over the years ($p<0.001$).

CONCLUSION: The findings underscore the ongoing inadequacy of interobserver agreement in the Gleason scoring system. Improvement suggestions involve conducting studies to ascertain in-clinic interobserver agreement enhancing training, facilitating information sharing, and employing accessible and easily applicable artificial intelligence-supported programs.

Keywords: Gleason score; interobserver agreement; prostate adenocarcinoma.

Cite this article as: *Cinar I, Cinar E. Assessment of interobserver variability in Gleason grading for prostate carcinoma. North Clin Istanb 2025;12(3):337–343.*

Gleason Score is the most widely used grading system for prostate adenocarcinoma and is the strongest predictor of the patient's clinical outcome. It is used in patient-specific treatment planning to decide if surgery, radiotherapy, or hormonal therapy will be applied. The scoring system proposed by Gleason et al. [1] includes 5 patterns based on the structural features of the tumor. The two most dominant patterns are given in order. The Gleason score represents the sum of these two patterns, while the grade group, introduced in the 2016 WHO classification, is a five-tier system emphasizing the most common pattern [1].

The Gleason score plays a critical role in the management and treatment of prostate cancer and is an important tool for predicting the biological aggressiveness and prognosis of cancer. Gleason score serves as a fundamental factor in determining patients' risk groups and evaluating treatment options. Patients are categorized into very low, low, intermediate, high, and very high-risk groups based on clinical factors such as Gleason score

Received: August 10, 2024 *Revised:* January 03, 2025 *Accepted:* February 20, 2025 *Online:* June 23, 2025

Correspondence: Ilkay CINAR, MD. Giresun Universitesi Tip Fakultesi, Patoloji Anabilim Dali, Giresun, Turkiye.
Tel: +90 454 310 16 00 e-mail: a.ilkaycinar@gmail.com

Istanbul Provincial Directorate of Health - Available online at www.northclinist.com

and prostate-specific antigen (PSA)levels [2]. For example, patients with a Gleason score of 6 generally have a better prognosis, with a low risk of metastatic disease and prostate cancer mortality. Therefore, less aggressive treatment options may be preferred for these patients. Patients with a Gleason score of 7, while having a lower risk of biochemical failure, present uncertainty regarding the need for adjuvant treatment due to their risk level. Treatment plans for these patients may vary based on individual risk factors and the patient's overall health status. Adjuvant therapy (androgen deprivation and/or radiotherapy) should be considered. Patients with a Gleason score greater than 7 have a higher risk of mortality. Consequently, more aggressive treatment approaches, such as radical prostatectomy and/or adjuvant therapy (e.g., hormone therapy), are recommended for these patients. This risk classification allows for the personalization of treatment strategies and directly impacts patients' long-term outcomes [2–5].

Therefore, interobserver variability in Gleason scoring can have a significant impact on risk assessment and clinical management when considered on a patient-by-patient basis. Like other grading systems, the Gleason scoring system contains intraobserver and interobserver variabilities [6–8]. In previous studies, it was observed that variability in Gleason scoring is more common among general pathologists and often trends to the assignment of lower Gleason scores [8]. Interobserver differences continue to exist. There is wide data to suggest that pathologist training and experience may influence the degree of interobserver agreement in the assignment of Gleason scores of biopsy specimens. In light of these findings, the aim of our study is to determine the interobserver agreement of prostate adenocarcinoma within the Gleason grading system in our center and the contributing factors.

Highlight key points

- Gleason grading is a key prognostic factor in prostate adenocarcinoma and guides treatment decisions.
- Minimal interobserver agreement was found among three pathologists in Gleason pattern, total score, and grade group evaluations (k=0.285–0.313).
- Agreement improved significantly over time among pathologists with longer collaborative experience (p<0.001).
- To improve diagnostic concordance, the implementation of regular training programs, intra-clinical case discussions, and the integration of AI-assisted technologies into pathology practice is recommended.

MATERIALS AND METHODS

This study has a retrospective design and was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the Giresun Training and Research Hospital Ethics Committee on July 3, 2023 (approval number: 19.06.2023/13).

Biopsy samples obtained through transrectal ultrasound-guided prostatic biopsy (TRUS-B) and diagnosed with prostatic adenocarcinoma between 2019 and 2023 were retrospectively retrieved through archive screening. Histomorphological evaluation of hematoxylin-eosin (HE) preparations was conducted using specimens prepared in accordance with our laboratory's routine procedure: fixed in 10% buffered formaldehyde, embedded in paraffin, cut at 5 μm thickness, and stained with HE. These evaluations were performed under a light microscope (Nikon Eclipse-Ci).

Samples with artifacts, extensive necrosis, or low tumor content were excluded from the study. Following these criteria, a total of 119 cases were included.

Histopathological Evaluation

Three general pathologists independently reevaluated all cases while being "blinded" to each other's assessments and previous reports. The pathologists recorded Gleason patterns, total Gleason scores, and grade groups. All samples were evaluated by all participants without considering any additional variables such as tumor size and location. All participating pathologists were experienced professionals with 7–10 years of practice in general pathology following their specialty training.

Determination of Gleason score

According to the Gleason grading system, the most common and the second most common patterns in the tumor were identified in order, and the Gleason score was obtained by adding the two values.

Determination of Grade group

Grade group was determined according to the criteria below.

Grade Group 1: Gleason score 6 (3+3)

Grade Group 2: Gleason score 7 (3+4)

Grade Group 3: Gleason score 7 (4+3)

Table 1. Kappa analysis reference values [9]

Value of Kappa	Level of agreement	% of data that are reliable
0–0.20	None	0–4
0.21–0.39	Minimal	4–15
0.40–0.50	Weak	15–35
0.60,0.79	Moderate	35–63
0.80–0.90	Strong	64–81
Above 0.90	Almost perfect	82–100

Grade Group 4: Gleason score 8 (4+4)

Grade Group 5: Gleason score 9 or 10 (5+4) or (4+5) or (5+5)

Statistics

Fleiss's kappa analysis was used to assess inter-observer agreement, and Spearman correlation analysis was used to evaluate the relationship between agreement and years. Normality tests were not applied given that the variables are categorical data. IBM SPSS Statistics 21.0 (IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.) and MS-Excel 2007 programs were used for statistical analyses and calculations. Statistical significance level was accepted as $p<0.05$. McHugh's [9] kappa analysis was performed in accordance with reference values (Table 1).

RESULTS

The analysis of inter-observer agreement among the three pathologists in determining Gleason patterns revealed minimal agreement (k=0.285). Specifically, weak agreement was observed for the patterns "3+3" and "3+5," while minimal agreement was noted for the patterns "4+4" and "4+5." No agreement was found for the pattern "5+5" ($p<0.001$). The detailed Gleason score agreement analysis is presented in Table 2.

Table 2. Inter-observer agreement for Gleason pattern assessment

Gleason pattern	Kappa	p	Agreement level
(3+3)	0.554	0.000	Weak
(3+4)	0.069	0.197	None
(4+3)	0.006	0.913	None
(4+4)	0.305	0.000	Minimal
(4+5)	0.221	0.000	Minimal
(5+4)	0.026	0.622	None
(5+5)	0.177	0.001	None
(3+5)	0.497	0.000	Weak
(5+3)	0.009	0.872	None

K: Fleiss' Kappa analysis, $p<0.005$.

For the evaluation of the Gleason score, minimal inter-observer agreement was also identified (k=0.309). Although the overall agreement across all scores was generally low, relatively better agreement was observed for score 6 (k=0.56). Minimal agreement was noted for Gleason scores "8," "9," and "10" ($p<0.001$). The agreement analysis for the total Gleason score from the three observers is summarized in Table 3.

When evaluated based on grade groups, minimal inter-observer agreement was again observed (k=0.313). Weak agreement was determined for Grade Group 1, while minimal agreement was noted for Grade Groups 4 and 5 ($p<0.001$). The inter-observer agreement analysis for grade groups is presented in Table 3.

Correlation in inter-observer agreement over the years was analyzed using the Spearman correlation test. The relationship between observers regarding the Glea-

Table 3. Kappa values for inter-observer agreement in Gleason scores and grade groups

Gleason score	Kappa	p	Agreement level	Grade group	Kappa	p	Agreement level
6	0.565	0.000	Weak	1	0.567	0.000	Weak
7	0.037	0.488	Minimal	2	0.069	0.197	None
8	0.297	0.000	Minimal	3	0.006	0.913	None

K: Fleiss' Kappa analysis, $p<0.005$.

Table 4. Inter-observer agreement for Gleason score relationship over years

Gleason score	Years	Observer-1	Observer-2	Observer-3
Observer 1	r	1.000	0.696	0.624
	p	0.000	0.000	0.000
Observer 2	r	0.696	1.000	0.705
	p	0.000	0.000	0.000
Observer 3	r	0.624	0.705	1.000
	p	0.000	0.000	0.000

r: Spearman correlation test.

Table 5. Inter-observer agreement Grade groups relationship over years

Grade groups	Years	Observer-1	Observer-2	Observer-3
Observer 1	r	1.000	0.736	0.626
	p	0.000	0.000	0.000
Observer 2	r	0.736	1.000	0.764
	p	0.000	0.000	0.000
Observer 3	r	0.626	0.764	1.000
	p	0.000	0.000	0.000

r: Spearman correlation test.

son score over the years is presented in Table 4, showing a moderate positive correlation between Observer 1 and Observer 2, Observer 1 and Observer 3, and a high positive correlation between Observer 2 and Observer 3, all statistically significant ($p<0.001$).

The relationship between observers for Grade Group over the years is depicted in Table 5, demonstrating a high positive correlation between Observer 1 and Observer 2, a high positive correlation between Observer 2 and Observer 3, and a moderate positive correlation between Observer 1 and Observer 3, all statistically significant ($p<0.001$).

DISCUSSION

Numerous grading systems are used in pathology practice, which inherently contain subjectivity. Previous studies have highlighted issues with inter-observer agreement [10]. Gleason scoring involves the identification of predetermined histopathological patterns of the tumor. While these pattern features have been previously described and may be well-known to observers, they remain subjective. Previous studies on inter-observer agreement in Gleason scoring have reported variable results [7, 8, 11–13], even among experienced uropathologists (k=0.43-0.68) [12]. In prostatic adenocarcinoma, tumor heterogeneity and the presence of morphologically borderline tumors may complicate Gleason scoring, as the rules for selecting which pattern to assign to the tumor are partly based on interpretation.

Our study results indicate minimal or weak overall agreement in Gleason grading, both in pattern determination and in Gleason scores, as well as Grade Groups. There was higher agreement in identifying lower grades compared to higher grades, with a decrease in agreement as the grade increased. The highest agreement was observed for a score of 6. Agreement was higher for pattern 3 compared to patterns 4 and 5, suggesting better consensus in recognizing well-formed and well-differentiated glands. However, Valdez and So [6] identified the best agreement at score 9 in their study, attributing it to the clearer identification of invasive neoplastic cells, layers, cords, solid nests, and necrosis, as well as greater tumor volume.

Our results also showed less agreement for scores 8 and 9, which may be attributed to challenges in evaluating subtle differences between poorly shaped and well-shaped glands, cribriform glands, and the difficulties in distinguishing between patterns 4 and 5 due to technical artifacts.

In the study by Agosti and Munari [14], it is emphasized that interobserver agreement is a significant issue in prostate cancer grading, with high variability among observers in evaluating Gleason score and grade groups. Notably, the agreement is particularly low when defining and quantifying Gleason pattern 4. For instance, the average agreement rate for Gleason scores among pathologists was found to be 71.4% for Gleason 3+3=6 and 56.4% for Gleason 3+4=7. Pattern 4 is identified as the most challenging pattern to evaluate, with different morphological types (such as malformed glands, fused glands, glomeruloid, and cribriform) associated with varying levels of observer agreement [14]. In our study, patterns as high as 4 and 5 were observed, with compliance found to be lower.

In the study by Ahsan et al. [7], the kappa value for Gleason patterns between two pathologists was reported to be 0.556, indicating a moderate level of agreement. The

highest agreement (23.4%) was observed for Gleason score 7, while the overall agreement rate was found to be 67%. These findings suggest that the level of agreement among observers was relatively higher in this study [7].

Additionally, the research conducted by Van der Slot et al. [15] involved six pathologists evaluating 80 radical prostatectomy specimens to examine observer variability. They assessed parameters such as the percentage of Gleason patterns 4 and 5, as well as the presence of invasive cribriform and intraductal carcinoma. It was noted that the agreement on Gleason pattern 4 and the presence of invasive cribriform and intraductal carcinoma was only moderate [15].

In recent years, research findings have suggested that tumors with a Gleason score of 6 may not meet the established criteria for classification as cancer. Ultimately, it is recommended that these tumors should not be referred to as cancer, aiming to prevent patients from undergoing unnecessary treatments and minimizing the psychological impact associated with the cancer label [16]. From our perspective, maintaining a high level of interobserver agreement is crucial, especially when assessing pattern 3 tumors. Modifying a patient's treatment plan based on interobserver variability can lead to significant consequences. In our study, despite observing a relatively higher agreement in score 6 tumors compared to other patterns, it remained weak (k=0.56).

This study also revealed a progressive improvement in interobserver agreement over time. This positive trend is likely attributed to increased consultation, information sharing, and the exchange of perspectives among observers. Similarly, a prior study by Mikami et al. [17] utilized a training atlas to assess the impact of training on the Gleason grading system. The results indicated a more favorable interobserver variability in the group that underwent training. Observers reported that training notably enhanced compliance, particularly in lower-grade assessments [17]. Allsbrook et al. [8] identified that the most significant factor linked to improved interobserver agreement was acquiring knowledge of the Gleason grading through participation in meetings or courses.

In our clinic, there was generally poor agreement observed among the three pathologists. We believe that conducting interobserver agreement studies within pathology clinics, coupled with in-clinical discussions and learning sessions on the results, will contribute to the enhancement of prostate biopsy reporting and elevate compliance levels.

Advancements in artificial intelligence promise to impact the field of pathology as well. There is a growing consensus that artificial intelligence could prove beneficial in addressing interobserver compatibility issues within the discipline [18]. In their study, Marrón-Esquivel et al. [19] exceeded the interobserver kappa score of 0.6946 obtained from the pathologist team by using several CNN models and found that the application of deep learning in score calculation could support pathologists in the diagnostic process. It was suggested that this approach could provide a valuable second opinion or act as a triage tool, enabling experts to concentrate on more aggressive cases.

One of the most commonly used methods to assess interobserver agreement is through Cohen's kappa and Fleiss' kappa coefficients. The kappa coefficient quantifies the observed agreement while accounting for the agreement that may occur by chance. As seen in many studies, the kappa statistic was utilized in our research. However, it is important to note that when there is an imbalance among the classes in the dataset, the kappa coefficient may yield low values. Additionally, the kappa statistic can vary depending on the prevalence of the disease and the number of evaluation categories [20].

Conclusion

Despite previous publications on the subject, the interobserver agreement within the Gleason score system remains inadequate. Improving interobserver agreement is critical in the diagnosis and treatment processes of prostate cancer. Lack of agreement can lead to patients being misclassified into incorrect risk groups, resulting in erroneous treatment strategies.

To enhance interobserver agreement, the following recommendations can be considered:

Training programs and standard protocols: Continuous education programs should be organized for pathologists. These programs should provide information on the current applications of the Gleason scoring system, histopathological evaluation techniques, and new developments. Training should include both theoretical and practical applications.

Assistive technologies: The integration of technologies such as artificial intelligence and machine learning can assist in the analysis of pathological images. These systems can support the pathologist's decision-making process, thereby reducing inconsistencies in evaluations.

Consensus meetings: Regular consensus meetings should be held among pathologists. In these meetings, discussions on specific cases can facilitate the convergence of different opinions, helping to develop a common understanding.

Quality control programs: Quality control programs should be implemented in pathology laboratories. These programs should include regular re-evaluation of samples to assess and improve interobserver agreement.

Data sharing and feedback: Platforms that encourage data sharing among pathologists should be established. These platforms can facilitate receiving and providing feedback on pathological evaluations.

Interdisciplinary collaboration: Collaboration among different disciplines such as oncology, urology, and pathology should be increased. This can help develop a more holistic approach in the treatment processes of patients.

Research and development: More research should be conducted to enhance interobserver agreement. These studies should focus on identifying the factors that lead to discrepancies and developing strategies to address these issues.

We advocate for conducting studies to assess interobserver agreement within pathology clinics and reviewing factors contributing to disagreement, ultimately leading to a more precise evaluation. In conclusion, improving interobserver agreement will contribute to achieving more accurate and reliable results in prostate cancer management. Education, information sharing, and the integration of technological innovations are critical to achieving this goal. These efforts will ensure better outcomes for patients in their treatment processes and enhance the overall quality of prostate cancer care.

Ethics Committee Approval: The Giresun Training and Research Hospital Ethics Committee granted approval for this study (date: 03.07.2023, number: 19.06.2023/13).

Informed Consent: Written informed consents were obtained from patients who participated in this study.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

Use of AI for Writing Assistance: The author declared that artificial intelligence-supported technologies were not used in this study.

Authorship Contributions: Concept – IC; Design – IC; Supervision – IC; Fundings – IC, EC; Materials – IC, EC; Data collection and/or processing – EC; Analysis and/or interpretation – IC, EC; Literature review – IC; Writing – IC; Critical review – IC, EC.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA; Grading Committee. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40:244-52. [CrossRef]
2. Mohler J, Antonarakis ES, Armstrong AJ, D'Amico AV, Davis BJ, Dorff T, et al. Prostate Cancer, Version 2.2019, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2019;17:479-505. [CrossRef]
3. Swanson GP, Basler JW. Gleason grading revision and outcomes. Am J Clin Pathol 2021;155:711-7. [CrossRef]
4. Mohler JL, Antonarakis ES. NCCN Clinical Practice Guidelines in Oncology: prostate cancer. J Natl Compr Canc Netw 2021;19:133-62.
5. Okubo Y, Sato S, Terao H, Yamamoto Y, Suzuki A, Hasegawa C, et al. Review of the developing landscape of prostate biopsy and its roles in prostate cancer diagnosis and treatment. Arch Esp Urol 2023;76:633-42. [CrossRef]
6. Valdez AL, So J. Interobserver variability of Gleason score and completeness of histopathology report in prostatic adenocarcinoma in prostate needle biopsy specimens among general pathologists in a multi-institutional setting. Philippine Soc Pathol 2018;3:1-5. [CrossRef]
7. Ahsan F, Khan AA, Asif M, Aslam M, Hamayun S, Din HU. Inter-observer variability in Gleason scoring system for histological grading of adenocarcinoma prostate. Pak Armed Forces Med J 2022;72:327-30. [CrossRef]
8. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. Hum Pathol 2001;32:81-8. Erratum in: Hum Pathol 2001;32:1417. [CrossRef]
9. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82. [CrossRef]
10. Cinar I. Inter observer agreement of the modified ishak histological activity index in chronic viral hepatitis among pathologists trained in different centers. Selcuk Med J 2023;39:178-82. [CrossRef]
11. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bambirra EA. Interobserver agreement of gleason score and modified Gleason Score in needle biopsy and in surgical specimen of prostate cancer. Int Braz J Urol 2007;33:639-46. [CrossRef]
12. Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol 2001;32:74-80. [CrossRef]
13. Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. Scand J Urol 2016;50:420-4. [CrossRef]
14. Agosti V, Munari E. Current issues in prostate cancer histopathology. Asian J Androl 2024;26:575-81. [CrossRef]
15. Van der Slot MA, Hollemans E, den Bakker MA, Hoedemaeker R, Kliffen M, van Leenders GJLH. Inter-observer variability of cribriform architecture and percent Gleason pattern 4 in prostate cancer: relation to clinical outcome. Virchows Arch 2021;478:249-56. [CrossRef]
16. Tang Y, Hu X, Wang Y, Li X. Gleason score 6: overdiagnosis and overtreatment? Asian J Surg 2023;46:2637-8. [CrossRef]
17. Mikami Y, Manabe T, Epstein JI, Shiraishi T, Furusato M, Tsuzuki T, Matsuno Y SH. Accuracy of gleason grading by practicing pathologists

and the impact of education on improving agreement. Hum Pathol 2003;34:658-65. [CrossRef]
18. Şerbănescu MS, Manea NC, Streba L, Belciug S, Pleşea IE, Pirici I, et al. Automated Gleason grading of prostate cancer using transfer learning from general-purpose deep-learning networks. Rom J Morphol Embryol 2020;61:149-55. [CrossRef]
19. Marrón-Esquivel JM, Duran-Lopez L, Linares-Barranco A, Dominguez-Morales JP. A comparative study of the inter-observer variability on Gleason grading against Deep Learning-based approaches for prostate cancer. Comput Biol Med 2023;159:106856. [CrossRef]
20. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303-8. [CrossRef]