# Reproducibility of Gleason Grading of Prostatic Adenocarcinoma


## *A.T. Atanda, A.B. Umar, I. Yusuf and M.I. Imam*

*Department of Pathology, Bayero University/ Aminu Kano Teaching Hospital, Kano, Nigeria



**Abstract**

***Background:*** *Gleason grading system for carcinomas of the prostate is important in determining treatment and outcome for patients. However, there is need to audit its use among pathologists to ensure reproducibility, thus avoiding undesirable consequences of inappropriate treatment.*

***Materials and Methods:*** *Ten slides made from needle biopsies of varying primary patterns and scores were administered to 11 general pathologists. Their ratings were measured against consensus expert ratings of the lesions and degrees of inter- and intra-rater agreements were measured using kappa statistics.*

***Results:*** *The inter-rater agreement for primary pattern recognition showed a range of kappa from 0.07 to 0.47 with most raters (45.5%) showing fair agreement with consensus rating. Overall kappa for primary pattern was 0.25 (fair agreement). Pattern underrating occurred overall in 49.1% of ratings and overrating in 3.6% with Gleason pattern 4 being the most underrated. Kappa coefficient for intra-rater consistency ranged from 0.29 to 0.78 (fair to substantial) with intra-rater consistency being highest for Gleason pattern 3. The inter-rater agreement for Gleason scores showed a range of kappa from – 0.12 to 0.54 (poor to moderate) and majority of raters (54.5%) being in the slight agreement range of kappa. The overall kappa was 0.35 (fair reproducibility). Gleason score 7, was undergraded in 63.6% of ratings, score group 8 – 10 by 45.5% and group 5 – 6 was undergraded in 38.6% of ratings.*

***Conclusion:*** *the study shows fair inter- and intra-rater consistency in Gleason pattern recognition and scoring with underscoring being the major factor identified. This underscores the need for constant revision of the use of grading systems to ensure consistency among raters.*



**Keywords**: Gleason grade; pattern; score; undergrading; kappa; inter-rater


## Introduction

Cancer of the prostate, accounting for about 13.6% of all male cancers, is not only one of the most common cancers worldwide but also the most common male cancer in Nigeria,[1, 2] and like other tumours clinical staging and histological grading are very important in its management. Thus, for several years, much research effort was channeled into finding reproducible ways of grading the tumour. Eventually, in 1966, Donald F. Gleason developed the current grading system that is named after him.[3] The initial grading based on glandular architectural pattern was later refined


***Correspondence to:*** Dr A. T. Atanda, Department of Pathology, Aminu Kano Teaching Hospital, Kano, Nigeria. PMB 3452, Kano. Post code 700001 E-mail: akinzo123@gmail.com



**No conflicts of interest have been declared by the authors**

to a 5-patterned system.[4] Even though several alternatives and modifications of the original Gleason grading have been proposed, including that of Helpap[5] the Gleason grading has survived as a veritable tool for managing prostatic cancer.

The Gleason grading system is carried out at low-power microscope objective and assesses gland to gland proximity, degree to which tumour cells are able to form glands and gland architecture. This is graded 1 to 5. The predominant grade (primary pattern) together with the less predominant grade (secondary pattern) are added together to derive the Gleason score. A tertiary pattern is also considered especially in cases with a more heterogeneous morphology and a minor pattern of higher grade.[6]

The Gleason score has shown great value in influencing choice of therapy and for prognostication.[7] Gleason score of 7, for which adjuvant chemotherapy may be offered, has been regarded as the Rubicon for decision taking as regards options for therapy. With Gleason scores of 6 and lower most surgeons adopt the "watch and wait" strategy and for patients with score of 8 and higher chemotherapy and radiotherapy are usually considered. It has also proven to be an independent predictor of medical therapy failure.

The need to correctly rate the Gleason patterns and score adenocarcinomas of the prostate using this system can thus not be overemphasized. To this end this study is carried out to audit the degree of inter-rater and intra-rater agreement of the Gleason grading system among general pathologists to ensure it is being done as accurately as possible.

## Materials and Methods

Criteria similar to those of the 2005 International Society of Urology Pathologists (ISUP)[8] were used to assess 10 core needle biopsies of the prostate processed and stained with Haematoxylin and Eosin showing adenocarcinomas with uniform primary patterns from 2 to 5. The slides were reviewed and consensus patterns were assigned by the author and an expert in grading prostatic adenocarcinomas. The slides included one primary pattern 2 and one primary pattern 4 lesions and four each of primary patterns 3 and 5. The latter two were chosen for assessment of intra-observer agreement because of difficulties and controversies usually associated with pattern 3 and the relative ease of assessing pattern 5. To be assigned the primary pattern it had to be seen in >50% of the tumour for each case.

The 10 slides were then administered to 11 general pathologists. Their primary pattern ratings and Gleason scores for each slide were then compared with consensus ratings and results compared using kappa statistics (at 95% confidence interval and $p < 0.001$) to determine inter- and intra-observer agreement. Minitab version 15 statistical package was used for the computations.

Based on the calculated kappa coefficients the degrees of agreement were then evaluated as: <0 (poor reproducibility); 0.01 – 0.2 (slight reproducibility); 0.21 – 0.40(fair reproducibility); 0.41 – 0.60 (moderate reproducibility); 0.61 – 0.80 (substantial reproducibility); 0.81 – 0.99 (almost perfect reproducibility).

## Results

Eleven pathologists rated the ten slides in one hundred and ten readings. As shown in Table 1, pattern 2, assigned as primary pattern in 38 (34.5%) of 110 ratings was the most frequently assigned pattern, followed by pattern 3 in 32 (29.1%) of ratings done. The least common primary pattern assigned by the pathologists was pattern 1 (2.7% of ratings). Underrating occurred overall in 54 (49.1%) of ratings; overrating in 3.6% (4/110) and appropriate rating in 47.3% of instances.

The inter-rater reliability analysis using kappa statistics was performed to determine

**Table 1.** Distribution of ratings of primary Gleason pattern by raters relative to consensus

| Slide Number | Consensus Pattern | Gleason Patterns 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | 3 | 0 | 3 | 7 | 1 | 0 |
| 2 | 2 | 1 | 10 | 0 | 0 | 0 |
| 3 | 5 | 0 | 3 | 4 | 2 | 2 |
| 4 | 3 | 1 | 7 | 3 | 0 | 0 |
| 5 | 5 | 0 | 1 | 0 | 3 | 7 |
| 6 | 4 | 0 | 4 | 5 | 2 | 0 |
| 7 | 3 | 0 | 4 | 5 | 2 | 0 |
| 8 | 5 | 0 | 0 | 0 | 1 | 10 |
| 9 | 3 | 1 | 3 | 6 | 1 | 0 |
| 10 | 5 | 0 | 3 | 2 | 6 | 0 |
| **Total** | | **3** | **38** | **32** | **18** | **19** |

**Table 2.** Distribution of Gleason scores as graded by raters relative to consensus scores

| Slide No. | Consensus score | Gleason 2 – 4 | Gleason 5 – 6 | Gleason 7 | Gleason 8 – 10 |
|---|---|---|---|---|---|
| 1 | 8 | 2 | 2 | 4 | 3 |
| 2 | 5 | 8 | 3 | 0 | 0 |
| 3 | 8 | 1 | 1 | 5 | 4 |
| 4 | 5 | 6 | 5 | 0 | 0 |
| 5 | 10 | 0 | 1 | 1 | 9 |
| 6 | 9 | 3 | 6 | 2 | 0 |
| 7 | 5 | 1 | 6 | 3 | 1 |
| 8 | 10 | 0 | 0 | 1 | 10 |
| 9 | 6 | 2 | 9 | 0 | 0 |
| 10 | 7 | 1 | 6 | 1 | 3 |
| **Total** | | **24** | **39** | **17** | **30** |

consistency among raters relative to the consensus rating, and as shown in Table 3, the inter-rater agreement for the raters showed a range of kappa from 0.07 to 0.47. Two (18.1%) of the raters had slight levels of agreement with consensus patterns; 5 (45.5%) fair agreement; and 4 (36.4%) moderate agreement with consensus rating. The overall inter-rater consistency was 0.25 (fair agreement).

Intra-rater consistency for Gleason pattern 3 (Cohen kappa statistic) was mostly moderate (45.5%) with 3 raters showing substantial consistency. Only 2 (18.1%) of the 11 raters showed substantial intra-rater consistency for Gleason pattern 5. The majority (5/11; 45.5%) only showing fair intra-rater consistency.

Table 2, on the other hand, shows the pattern of grading (primary + secondary patterns) as rendered by the raters. Grades 5 – 6 accounted for 39 (36%) of the 110 ratings, followed in frequency by grades 8 – 10 with 30 ((27%) of the ratings and borderline, grade 7, accounting for 17 (15%) of the ratings. Grade 2 – 4 accounted for 22% of the ratings. Relative to the consensus Gleason scores for the 10 needle biopsies, raters undergraded the scores 51.8% of times and over graded only in 7.3% of instances. Appropriate scoring was done in 40.9% of gradings. The intermediate grade, Gleason score 7, was underrated in 63.6% of ratings, followed in magnitude by the poorly differentiated group (8 – 10) which was undergraded by 45.5% and, least of all, the well differentiated group (5 – 6) which was undergraded by raters in 38.6% of ratings.

The inter-rater reliability analysis using kappa statistics was also performed to determine consistency among raters relative to the consensus Gleason scores, and as shown in Table 3, the inter-rater agreement for the raters showed a range of kappa from – 0.12 to 0.54, with majority of raters (6 of 11; 54.5%) being in the slight agreement range of kappa; 2 (18.2%) each being in the range of poor and fair agreement respectively. Only one rater was in moderate agreement with consensus scores. The overall kappa statistic was 0.35 and is in the realm of fair reproducibility.

## Discussion

Histopathology has over time faced the challenge of how to render mostly qualitative data into quantifiable formats so as to limit the impact of subtle to pronounced shades of

**Table 3.** Distribution of inter-rater and intra-rater kappa values for primary Gleason pattern recognition and Gleason grading

| Rater | Inter-observer Kappa coefficient | Intra-observer Kappa coefficient For Gleason Pattern 3 | Intra-observer Kappa coefficient For Gleason Partten 5 | Inter-rater Kappa coefficient For Gleason scores |
|---|---|---|---|---|
| 1 | 0.07 | 0.29 | 0.29 | 0.01 |
| 2 | 0.47 | 0.78 | 0.55 | 0.30 |
| 3 | 0.23 | 0.54 | 0.29 | – 0.20 |
| 4 | 0.42 | 0.55 | 0.78 | 0.20 |
| 5 | 0.47 | 0.55 | 0.55 | 0.19 |
| 6 | 0.14 | 0.29 | 0.29 | 0.18 |
| 7 | 0.33 | 0.78 | 0.29 | – 0.12 |
| 8 | 0.31 | 0.55 | 0.55 | 0.40 |
| 9 | 0.47 | 0.78 | 0.29 | 0.06 |
| 10 | 0.31 | 0.55 | 0.55 | 0.54 |
| 11 | 0.36 | 0.29 | 0.78 | 0.20 |

subjectivity peculiar to qualitative data. This problem is no less apparent than in grading of tumours. While efforts have been made, with varying degrees of success, to provide objective parameters for achieving this, the influence of subjectivity can still not be completely eliminated. The quantitative grading of prostatic adenocarcinoma, like other malignancies, faces no less the problem of subjectivity.

Another dilemma faced by pathologists is the problem of determining the "correct" diagnosis (or grade) of a lesion. For this study, the authors adopted the method of consensus and expert assessment of patterns in determining the true Gleason pattern and score for each case. Similar methodology was adopted by other studies[9, 10] with the same theme.

In our study, primary Gleason pattern 3 was underrated in 43.2% of ratings, while Gonzalgo *et al*[11] recorded as much as 47% of pattern 3 were underrated. Gleason pattern 4 was underrated in as much as 81.8% of the ratings by our pathologists and by 24% in the latter study.[11] While consensus primary pattern 5 lesions were underrated in 56.8% of ratings, a comparable magnitude of 57.6% was reported by Fajardo[12] in the United States. Furthermore, even though our kappa range of 0.07 to 0.47 (slight to moderate inter-rater agreement) for primary Gleason pattern is lower than the –

0.32 to 0.92 reported in an Indian study[10], their finding of a mode of 35% fair agreement is however, similar to our observation that most of our raters (45.5%), were in fair agreement with the consensus primary Gleason pattern. Our kappa range of – 0.20 to 0.54 (poor to moderate inter-rater agreement) for Gleason scoring is higher than the 0.16 – 0.29 obtained by McLean *et al*[13] and the 0.148 – 0.328 by Djavan *et al*[14] in Austria, it is lower than those of others (0.47 – 0.64)[15] and (-0.11 – 0.82)[10] with similar studies. The overall kappa statistic of 0.35 obtained from our study (fair inter-rater agreement) is comparable to the 0.36 obtained in Brazil,[16] but lower than the moderate agreement (0.49 and 0.44) found for some Japanese and American general pathologists[17] and that by Goodman *et al*[18] of 0.61.

Our raters undergraded the scores overall in 51.8% of ratings, a finding as high as the 50.1 % reported for needle biopsy grading among pathologists in Austria.[14] Analysis from other studies[10, 17, 19, 20] have also identified undergrading as the major factor responsible for low inter-rater agreement. Undergrading of Gleason score group 7 (by 63.6%) and group 8 – 10 (by 45.5%) by our raters is comparably higher than the 47% and 47% obtained respectively for the two Gleason groups by Allsbrook.[15] While our study showed overgrading in only 7.3% of total ratings, values of 12.7% and 33.7% have been demonstrated by others.[10, 14]

Inter-rater agreement appears to reflect levels of experience and expertise in the assessment of the Gleason patterns; with general pathologists, like those who participated in our study, having been shown to be more likely to underscore.[19, 20] Our study included an exceptional case with consensus rating of primary Gleason pattern 2 which 91% (10 of 11) raters correctly identified. However, in 28.2% of instances raters included Gleason patterns 1 and 2 in their assessments where these patterns should not have been assigned without immunohistochemistry (IHC) backup. This may be a result of inadequate awareness of studies[8, 12] which have demonstrated poor correlation of these patterns with eventual radical prostatectomy scores. Such finding of poor correlation prompted the ISUP consensus on Gleason Grading[8] to recommend that scores 2 – 4 should only rarely, if ever, be assigned in needle biopsy settings. These, erstwhile widely reported grades/patterns, in the current light of IHC, would now more probably be diagnosed as atypical adenomatous hyperplasia.[21]

Gleason pattern 3, as next most commonly rated in our study, is recommended to be the least common pattern assigned in needle biopsies.[8] Yet in the context of needle biopsies great caution needs also be exercised in the rating of Gleason pattern 3, as cribriform intra-epithelial lesion is now recognized based on IHC evaluation of such lesions.[21] Only cribriform lesions in which IHC demonstrates loss of basal staining and or where extra-prostatic extension is demonstrable or perineural invasion is present should be codified as grade 3 adenocarcinomas. Additionally most urologists reports confirmed cribriform carcinoma as Gleason pattern 4 or higher[22] and this is supported by observations that this pattern has been associated with an aggressive course.[23] Recognition of cribriform architecture in pattern 5 is now also recognized, particularly when associated with comedonecrosis.[8] A review of current recommendations of ISUP 2005 and emerging observations from different studies hint at the reality that assessment of the Gleason grading system is bound for more complexity than Gleason originally described!

Though the performance of our participating pathologists in this audit activity is of varying levels relative to other pathologists the world over, a common denominator to all is the issue of undergrading, with the problem of pattern recognition being the major underlying factor. Similar also to other earlier referenced studies[10, 15, 17] the assessment of Gleason patterns 3 and 4, and by extension score 7 (3+4 = 7 and 4+3 = 7) is often problematic. An approach to

mitigating the problem has been the institution of regular tutorials including digital image reviews, subjection to external quality assurance schemes and not shying away from getting second opinions on necessary cases.[19, 24]

## Conclusion

In conclusion, this study shows there is fair inter- and intra-rater consistency in Gleason pattern recognition and scoring, with undergrading being the major factor identified. This emphasises the need for constant revision of the use of grading systems to ensure consistency among raters.

## References


1. GLOBOCAN 2008 database (version 1.2) [Internet]. –[cited 2013 feb 4]. Available from: http:// www.globocan.iarc.fr
2. Ogunbiyi JO and Shittu BO. Increased incidence of prostate cancer in Nigeria. J Natl Med Assoc 1999; 91: 159-164.
3. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep 1966; 50: 125-128.
4. Gleason DF and Wellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol 1974; 11: 58-64.
5. Mahlke U, Ulman A and Kunz J. [Prognostic significance of prostatic carcinoma grading according to Helpap]. Verh Dtsch Ges Pathol 1993; 77: 82-85.
6. Epstein JI. An update of the Gleason grading system. J Urol 2010; 183: 433-440.
7. Roehl KA, Han M, Ramos CG, Antenor JA and Catalona WJ. Cancer progression and survival rates following anatomical radical retropubic prostatectomy in 3,478 consecutive patients: long-term results. J Urol 2004; 172: 910-914.
8. Epstein JI, Allsbrook WC Jr, Amin MB and Egevad LL. ISUP Grading Committee. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol 2005; 29: 1228-1242.
9. Melia J, Moseley R, Ball RY, Griffiths DF, Harnden P, Jarmulowicz M, *et al.* A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology 2006; 48: 644-654.
10. Singh RV, Agashe SR, Gosavi AV and Sulhyan KR. Inter-observer reproducibility of Gleason grading of prostatic adenocarcinoma among general pathologists. Indian J Cancer 2011; 48: 488-495.
11. Gonzalgo ML, Bastian PJ, Mangold LA, Trock BJ, Epstein JI, Walsh PC, *et al.* Relationship between primary Gleason pattern on needle biopsy and clinic-pathologic outcomes among men with Gleason score 7 adenocarcinoma of the prostate. Urology 2006; 67: 115-119.
12. Fajardo DA, Miyamoto H, Miller JS, Lee TK and Epstein JI. Identification of Gleason pattern 5 on prostatic needle core biopsy: frequency of underdiagnosis and relation to morphology. Am J Surg Pathol 2011; 35: 1706-1711.
13. McLean M, Srigley J, Banerjee D, Warde P and Hao Y. Inter-observer variation in prostate cancer Gleason scoring: Are there implications for the design of clinical trials and treatment strategies? Clin Oncol (RColl Radiol) 1997; 9: 222-225.
14. Djavan B, Kadesky K, Klopukh B, Marberger M and Roehrborn CG. Gleason scores from prostate biopsies obtained with 18-gauge biopsy needles poorly predict Gleason scores of radical prostatectomy specimens. Eur Urol 1998; 33: 261-270.

15. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG and Amin MB. Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Hum Pathol 2001; 32: 74-80.
16. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scallon JD and Bambira EA. Inter-observer agreement of Gleason Score and Modified Gleason Score in Needle Biopsy and Surgical specimen of Prostate Cancer. Int J Braz Urol 2007; 33: 639 – 651.
17. Oyama T, Allsbrook WC Jr, Kurokawa K, Matsuda H, Segawa A, Sano T, Suzuki K and Epstein JI. A comparison of inter-observer reproducibility of Gleason grading of prostatic carcinoma in Japan and the United States. Arch Pathl Lab Med 2005 Aug; 129: 1004-1010.
18. Goodman M., Ward K C, Osunkoya AO, Datta MW, Luthringer D, Young AN, *et al*. Frequency and determinants of disagreement and error in Gleason scores: A population-based study of prostate cancer. Prostate 2012; 72: 1389–1398.
19. Griffiths DF, Melia J, McWilliam LJ, Ball RY, Grigor K, Harnden P, *et al*. A study of Gleason score interpretation in different groups of UK Pathologists; techniques for improving reproducibility. Histopathology 2006; 48: 655-662.
20. Renshaw AA, Schultz D, Cote K, Loffredo M, Ziemba DE and D'Amico AV. Accurate Gleason grading of prostatic adenocarcinoma in prostate needle biopsies by general pathologists. Arch Pathol Lab Med 2003; 127: 1007-1008.
21. Epstein JI. Gleason score 2 – 4 adenocarcinoma of prostate on needle biopsy: a diagnosis that should not be made. Am J Surg Pathol 2000; 24: 477-478.
22. Shah RB. Current perspectives on the Gleason grading of prostate Cancer. Arch Pathol Lab Med 2009; 133: 1810-1816.
23. Latour M, Amin MB, Billis A, Egevad L, Grignon DJ, Humphrey PA, *et al.* Grading of invasive cribriform carcinoma on prostate needle biopsy: an inter-observer study among experts in genitourinary pathology. Am J Surg Pathol 2008; 32: 1532-1539.
24. Harnden P, Coleman D, Moss S, Kodikara S, Griffin NR and Melia J. Evaluation of the use of digital images for a national prostate core external quality assurance scheme. Histopathology 2011; 59: 703-709.