Inter/Intra-Observer Reproducibility of Gleason Scoring in Prostate Adenocarcinoma in Iranian Pathologists

Alireza Abdollahi,[1] Alipasha Meysamie,[2] Sara Sheikhbahaei,[2] Ali Ahmadi,[1] Hedieh Moradi-Tabriz,[1] Mohammadreza Bakhshandeh,[1] Hassan Hosseinzadeh[1]

[1]Department of Pathology, Tehran University of Medical Sciences, Tehran, Iran
[2]Department of Community Medicine, Tehran University of Medical Sciences, Tehran, Iran

Corresponding Author:

Alireza Abdollahi, MD
Department of Pathology, Imam Khomeini Teaching Hospital, Tehran University of Medical Sciences, Tehran, Iran

Tel: +98 912 122 0588
Fax: +98 21 8826 9844
E-mail: dr_p_abdollahi@yahoo.com

Received December 2010
Accepted April 2011

Purpose: To measure the level of inter/intra-observer reproducibility among pathologists as far as Gleason scoring of adenocarcinoma of the prostate is concerned.

Materials and Methods: A total of 101 prostate biopsy slides, diagnosed with adenocarcinoma of the prostate by five pathologists from different education centers, were exposed to Gleason scoring. Two months later, the slides were re-examined by three of the same pathologists. Thereafter, the kappa was calculated for the data provided in the first and second reports of each pathologist and compared between pathologists.

Results: Inter-observer reproducibility was inappropriate, but intra-observer diagnostic reproducibility was almost perfect with a corresponding percentage of agreement of 85.2%.

Conclusion: The inter-observer reproducibility was poor.

Keywords: prostatic neoplasms, neoplasm grading, methods, humans

INTRODUCTION

Prostate cancer (PCa), besides the skin cancer, is the most prevalent type of cancer in men in the United States. It is also the second leading cause of cancer-related deaths in men, just following lung cancer.[1] The overall prostate cancer detection rate in our community is 3.5%.[2,3] The gold standard in the diagnosis of the PCa is biopsy and making a histological diagnosis of carcinoma.[1,4,5] When the tissue sample indicates presence of carcinoma, its Gleason scoring is one of the most important elements in reporting. In this method, tumors are graded, based on their pattern of growth and the level of differentiation, from 1 to 5; grade 1 has the lowest and grade 5 the highest level of differentiation.

One of the contributing factors to this observed upgrading in Gleason scoring is the level of pathologist experience.[6,7] Since Gleason score is one of the most important prognostic factors for the outcome of treatment in PCa and even determines the treatment of choice for the tumor,[8-12] a high degree of precision in its reporting and the agreement among the reports of the different pathologists for the same sample are crucial issues. Despite the fact that Gleason scoring is simple, there is an inter-observer variability of the scores. Gleason once said that "If I re-score my previously scored samples, in 50% of cases, I report the same scores and in 85.2%, ± 1 standard deviation of the previous scores".[11]

In previous studies, average kappa (K) that indicates the concordance rate of the report varied between 0.16 and 0.836.[13-15] For example, in a study conducted by Rodriguez-Urrego and colleagues, the inter-observer agreement was excellent with k = 0.72 and the intra-observer agreement was very good with k = 0.65 and even more.[16]

In this study, regarding the crucial importance of the reproducible and concordant reporting of the samples among different pathologists, we want to obtain approximations for these two variables among the pathologists working in Iran.

MATERIALS AND METHODS

In this cross-sectional study, 101 tissue samples of the prostate adenocarcinoma obtained through needle biopsy were re-examined to be Gleason-scored.

First, the paraffin-embedded tissues were cut into thin microscopic sections, and after being stained with Hematoxylin and Eosin, they were sent to five

Table 1. Gleason's microscopic grading system of the prostate carcinoma.

Stage	Description
1	Single, separate, uniform glands in closely packed masses with a definite, usually rounded edge limiting area of tumor.
2	Single, separate, slightly less uniform glands, loosely packed (separated by small amounts of stroma), with less sharp edge.
3a	Single, separate, much more variable glands, may be closely packed, but usually irregularly separated, with ragged, poorly defined edge.
3b	Like 3a, but very small glands or tiny cell clusters.
3c	Sharply and smoothly circumscribed rounded masses of papillary or loose cribriform tumor (papillary intraductal tumor).
4a	Raggedly outlined, raggedly infiltrating, fused glandular tumor.
4b	Like 4a, with large pale cells (hypernephroid).
5a	Sharply circumscribed, rounded masses of almost solid cribriform tumor, usually with central necrosis (comedocarcinoma).
5b	Ragged masses of anaplastic carcinoma with only enough gland formation or vacuoles to identify it as adenocarcinoma.

randomly selected pathologists to be scored using the Gleason scoring system (before modification by ISUP, 2005). The microscopic grading system was based on the degree of glandular differentiation and the growth pattern of the tumor compared with the stroma (Table 1).

Sections that cannot be scored, those extracted from patients previously treated with anti-androgenic drugs or radiotherapy, and samples containing less than 5 malignant acini were excluded from the study.

After selection of the samples, a code was given to each of them. Thereafter, the scores given by each of the five pathologists were recorded in a data sheet. Two months later, the same samples with altered code were sent back to three of the pathologists to be re-scored. Finally, the concordance rate was measured among the five pathologists.

This research was carried out according to the principles of the Declaration of Helsinki. The local Ethics Medical Committee of Tehran University of Medical Sciences approved the study protocol.

Our statistical analysis included calculation of kappa for each pathologist based on his first and second data report and comparison of kappa between pathologists. Kappa varied between 0 and 1; the greater the kappa, the higher the concordance rate. Kappa value of 0 to 0.20 indicated slight

Table 2. Percentages of agreement and Kappa values of all possible pair combination of 5 pathologists' grading scores.*†

	O1T1	O2T1	O3T1	O4T1	O5T1	O3T2	O4T2	O5T2	
		38.30% (28.77% to 47.83%)	40.00% (30.40% to 49.60%)	36.50% (27.06% to 45.94%)	60.00% (50.40% to 69.60%)	35.70% (26.31% to 45.09%)	39.10% (29.54% to 48.66%)	59.10% (49.46% to 68.74%)	Agreement, %
O1T1		0.24	0.25	0.19	0.48	0.19	0.21	0.47	Kappa
			46.60% (36.82% to 56.38%)	34.50% (25.18% to 43.82%)	50.00% (40.20% to 59.80%)	41.40% (31.75% to 51.05%)	31.00% (21.94% to 40.06%)	46.60% (36.82% to 56.38%)	Agreement, %
O2T1			0.34	0.19	0.38	0.28	0.15	0.35	Kappa
				37.10% (27.63% to 46.57%)	52.60% (42.81% to 62.39%)	87.90% (81.51% to 94.29%)	31.00% (21.94% to 40.06%)	48.30% (38.51% to 58.09%)	Agreement, %
O3T1				0.19	0.4	0.85	0.12	0.34	Kappa
					41.00% (31.36% to 50.64%)	37.60% (28.11% to 47.09%)	78.60% (70.56% to 86.64%)	40.20% (30.59% to 49.81%)	Agreement, %
O4T1					0.25	0.2	0.72	0.24	Kappa
						52.10% (42.31% to 61.89%)	38.50% (28.96% to 48.04%)	88.90% (82.74% to 95.06%)	Agreement, %
O5T1						0.39	0.21	0.86	Kappa
							30.80% (21.75% to 39.85%)	48.70% (38.90% to 58.50%)	Agreement, %
O3T2							0.11	0.35	Kappa
								41.00% (31.36% to 50.64%)	Agreement, %
O4T2								0.24	Kappa

* $P < .001$.

† O1 indicates observer 1; O2, observer 2; O3, observer 3; O4, observer 4; O5, observer 5; T1, Time 1; and T2, Time 2.

agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and ≥ 0.81 was regarded as almost perfect agreement.[15] Eventually, the data were analyzed both descriptively and analytically using SPSS (the Statistical Package for the Social Sciences, Version 15.0, SPSS Inc, Chicago, Illinois, USA) and STATA 8 softwares.

RESULTS

Percentages of agreement and Kappa values of all possible pair combination of scores of five pathologists are shown in Table 2.

Overall kappa values in different Gleason scores for the five observers were calculated (Table 3). The mean kappa value was 0.29 (fair agreement). Using weighted kappa values, there was no significant difference in inter-observer agreement between poorly differentiated and moderately differentiated tumors. Kappa values of scores 4 and 5 were not significant.

Intra-observer diagnostic reproducibility was almost perfect with a corresponding percentage of agreement of 85.2%.

Table 3. Inter-observer reproducibility of Gleason's grading system for prostatic carcinoma.*

Gleason Score	Kappa	Prob > Z
4	- 0.0070	0.5939
5	0.0417	0.0788
6	0.4033	< 0.001
7	0.1855	< 0.001
8	0.2318	< 0.001
9	0.3402	< 0.001
10	0.3964	< 0.001
Combined	0.2896	< 0.001

*There were 5 observers per subject.
Scores 4 and 5 were not significant.

DISCUSSION

Today, the Gleason system (prostate adenocarcinoma grading system) is widely used for tumor grading, crucial for both patients and doctors.

We found an extremely low reproducibility between contributing pathologists (0.29 = fair agreement). However, two months later, the reproducibility has a good level when the slides were reported for the second time (intra-observer) (85.2% = perfect agreement). This result can show that Iran still lacks an integrated and regular education system in pathology.

In addition, there are obvious limitations in the accuracy of grading based on the small amount of tissue available from needle biopsies of the prostate.

On the other hand, one should recognize a pathological misinterpretation. Differences of opinion are related to different interpretations of tumor grading, which is a qualitative indicator. Naturally, this qualitative factor may be interpreted differently by pathologists. Therefore, the difference in interpretation should not be construed as an error.

In this study, the reproducibility was meaningfully proportionate with Gleason score of the samples. It seems that the reproducibility would rise when the score goes from 2 to 10.

Ozdamar and colleagues reported an acceptable inter-observer variation for Gleason-style grading.[17] Furthermore, Allsbrook and associates examined 46 cases of cancer for grading. Ten pathologists were involved. The reproducibility stood at an acceptable level.[12] The reason behind different conclusions from these studies might be related to the pathology education system. Holding meetings for exchange of views, conferences, journals, and group studies may bring views and interpretations together. The lack of such programs in our country must explain the differences.

The study by Mulay and coworkers on 40 patients with cancer produced an inter-observer reproducibility between 0.36 and 0.64. After a web-based training course for pathologists contributing to this project, the indicator soared significantly. This study is well indicative of the significance of regular training to pathologists.[18]

This study faced restrictions, such as undermanned samples. Therefore, studies with more pathologists and samples are recommended for the future.

CONCLUSION

Given the significance of grading in the prostate carcinoma, regular and effective training courses are strongly recommended for pathologists in order to raise intra/inter-observer reproducibility.

ACKNOWLEDGEMENTS

The authors are sincerely grateful to the Vice Chancellor for Research Affairs at Tehran University of Medical Sciences for his financial contribution to this study.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974;111:58-64.
2. Safarinejad MR. Population-based screening for prostate cancer by measuring free and total serum prostate-specific antigen in Iran. Ann Oncol. 2006;17:1166-71.
3. Hosseini SY, Moharramzadeh M, Ghadian AR, Hooshyar H, Lashay AR, Safarinejad MR. Population-based screening for prostate cancer by measuring total serum prostate-specific antigen in Iran. Int J Urol. 2007;14:406-11.
4. Bostwick DG. Gleason grading of prostatic needle biopsies. Correlation with grade in 316 matched prostatectomies. Am J Surg Pathol. 1994;18:796-803.
5. Vira MA, Tomaszewski JE, Hwang WT, et al. Impact of the percentage of positive biopsy cores on the further stratification of primary grade 3 and grade 4 Gleason score 7 tumors in radical prostatectomy patients. Urology. 2005;66:1015-9.
6. D'Amico AV, Renshaw AA, Arsenault L, Schultz D, Richie JP. Clinical predictors of upgrading to Gleason grade 4 or 5 disease at radical prostatectomy: potential implications for patient selection for radiation and androgen suppression therapy. Int J Radiat Oncol Biol Phys. 1999;45:841-6.
7. Kulkarni GS, Lockwood G, Evans A, et al. Clinical predictors of Gleason score upgrading: implications for patients considering watchful waiting, active surveillance, or brachytherapy. Cancer. 2007;109:2432-8.
8. Patel AA, Chen MH, Renshaw AA, D'Amico AV. PSA failure following definitive treatment of prostate cancer having biopsy Gleason score 7 with tertiary grade 5. JAMA. 2007;298:1533-8.
9. Whittemore DE, Hick EJ, Carter MR, Moul JW, Miranda-Sousa AJ, Sexton WJ. Significance of tertiary Gleason pattern 5 in Gleason score 7 radical prostatectomy specimens. J Urol. 2008;179:516-22; discussion 22.
10. Albertsen PC, Hanley JA, Gleason DF, Barry MJ. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. JAMA. 1998;280:975-80.
11. Gleason DF. Histologic grading of prostate cancer: a perspective. Hum Pathol. 1992;23:273-9.
12. 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.
13. Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology. 2006;48:644-54.
14. Lotan TL, Epstein JI. Clinical implications of changing definitions within the Gleason grading system. Nat Rev Urol. 2010;7:136-42.
15. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ. 1992;304:1491-4.
16. Rodriguez-Urrego PA, Cronin AM, Al-Ahmadie HA, et al. Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. Hum Pathol. 2011;42:68-74.
17. Ozdamar SO, Sarikaya S, Yildiz L, Atilla MK, Kandemir B, Yildiz S. Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. Int Urol Nephrol. 1996;28:73-7.
18. Mulay K, Swain M, Jaiman S, Gowrishankar S. Gleason scoring of prostatic carcinoma: impact of a web-based tutorial on inter- and intra-observer variability. Indian J Pathol Microbiol. 2008;51:22-5.